SANDIA REPORT

SAND2017-9960
Unlimited Release
Printed October 2017

Robust Approaches to Quantification of Margin and Uncertainty for Sparse Data

Lauren Hund, Ben Schroeder, Kellin Rumsey, Nicole Murchison

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology and Engineering Solutions of Sandia, LLC.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online ordering:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2017-9960
Unlimited Release
Printed October 2017

Robust Approaches to Quantification of Margin and Uncertainty for Sparse Data

Lauren B. Hund
Statistical Sciences Group

Ben Schroeder
V&V, UQ, Credibility Processes

Kellin Rumsey
Statistical Sciences Group

Nicole Murchison
Human Factors Group

Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-9999

Abstract

Characterizing the tails of probability distributions plays a key role in quantification of margins and uncertainties (QMU), where the goal is characterization of low probability, high consequence events based on continuous measures of performance. When data are collected using physical experimentation, probability distributions are typically fit using statistical methods based on the collected data, and these parametric distributional assumptions are often used to extrapolate about the extreme tail behavior of the underlying probability distribution. In this project, we characterize the risk associated with such tail extrapolation. Specifically, we conducted a scaling study to demonstrate the large magnitude of the risk; then, we developed new methods for communicating risk associated with tail extrapolation from unvalidated statistical models; lastly, we proposed a Bayesian data-integration framework to mitigate tail extrapolation risk through integrating additional information. We conclude that decision-making using QMU is a complex process that cannot be achieved using statistical analyses alone.

Contents

1 **Introduction** **13**

1.1 Background on QMU for PhysSim at SNL 13

1.2 Perceptions of statistics in PhysSim QMU 17

1.3 Objectives of report 21

2 **Quantify model form risk** **23**

2.1 Methods 23

2.2 Results 25

2.2.1 Correct model form assumption: Normal distribution 25

2.2.2 Incorrect model form assumption: Tail sub-population 26

2.2.3 Incorrect model form assumption: T_5 distribution 28

2.3 Discussion 30

3 **Define model validation metrics** **35**

3.1 Methods 35

3.1.1 Percentile and tolerance interval estimation 35

3.1.2 Motivating example: QMU for launch safety device 38

3.1.3 New metrics 40

3.1.4 Degree of extrapolation 41

3.1.5 Model fit in the tails 42

3.1.6 Sensitivity to modeling assumptions 43

Performance of model sensitivity metric 45

3.2 Results 46

3.3 Discussion 49

4 Propose information integration models 53

4.1 Methods 53

4.2 Results 54

4.3 Discussion 57

5 Anctipated Impact 59

6 Conclusions 61

References 69

List of Figures

1.1 Definition of margin and uncertainty in QMU PhysSim process. 15

1.2 Histogram for the $n = 10$ closing time measurements, with the performance requirement in red. 15

1.3 Probability plot for $n = 10$ closing time measurements; Anderson-Darling $p = .38$. 16

1.4 Fitted distribution, 99.9^{th} percentile estimate (Q), 95% tolerance bound (TB), and performance requirement (PR) for the $n = 10$ observations. 17

1.5 True data generating model with vertical line at the true 99.9^{th} percentile (black) and the CD requirement (red). 18

2.1 Type 1 error rate scaling of Anderson-Darling test for normality applied to data from standard normal distribution at a 0.05 p-value 25

2.2 Scaling of mean performance metrics across sample size for 99.9% coverage tolerance intervals (assuming normality) for data from a standard normal distribution. 26

2.3 Scaling of mean performance metrics across sample size for 95% confidence tolerance intervals (assuming normality) for data from a standard normal distribution. 27

2.4 Distribution with right tail subpopulation (red solid line) and 99.9 percentile (red dashed lines) compared with 99.9 percentile based on equivalent normal distribution. 27

2.5 How the rejection rate of the AD test at a 0.05 p-value, scales with number of sample points for random combinations of datasets taken from the non-normal population shown in Figure 2.4. 28

2.6 Observed reliability (left plot) and percentile discrepancy (right plot) of the normal 99.9% coverage tolerance intervals applied to data from the distribution with a subpopulation on the right tail (Figure 2.4). 29

2.7 Observed reliability (left plot) and percentile discrepancy (right plot) of the normal 95% confidence tolerance intervals applied to data from the distribution with a subpopulation on the right tail (Figure 2.4). 30

2.8 T_5 distribution (red line) compared to normal distribution equivalent (green line). 99.9 percentiles of the distributions are shown as dashed lines. 31

2.9 How the rejection rate of the AD test at a 0.05 p-value, scales with number of sample points for random combinations of samples taken from non-normal population shown in Figure 2.8. 31

2.10 Observed reliability (left plot) and percentile discrepancy (right plot) of the normal 99.9% coverage tolerance intervals applied to data from a T_5 distribution. 32

2.11 Observed reliability (left plot) and percentile discrepancy (right plot) of the normal 95% confidence tolerance intervals applied to data from a T distribution with 5 degrees of freedom. 33

3.1 Illustration of a percentile estimate $\hat{Q}_r$ and a one-sided 95% confidence bound on the percentile estimate $\hat{Q}_{r,.95}$. 38

3.2 Histograms of the two simulated datasets, $n = 100$. 39

3.3 Probability plots for hot CT (left) and cold CT (right). 39

3.4 Two parametric distributions used as examples. 41

3.5 The non-parametric required sample size divided by the true sample size n as a function of the percentile of interest r at the 90% confidence level (dashed line) and 95% confidence level (solid line). 42

3.6 Simple example comparing a return-level plot (left) to a QQ plot (right) for tail extrapolation. 43

3.7 Posterior distributions for Q^1 under parametric model (orange) and Q^2 under more robust model (purple) for a 99.5^{th} percentile estimate based on $n = 100$ observations when the model is correct (left) and incorrect (right). 46

3.8 The probability of p_ε exceeding .9 as a function of ε when: (left) the normal model is correct and (right) the t_5 model is correct. 47

3.9 Return-level plot for hot CT (left) and cold CT (right). 48

3.10 Kernel density estimates of the posterior distributions for the 99.5^{th} percentile. . . . 49

3.11 Validation metrics. 50

4.1 Bayesian network representations of Bayesian hierarchical models used to determine tolerance intervals. 54

4.2 How the empirical reliability (left plot) and percentile discrepancy (right plot) of the 99.9% coverage Bayesian tolerance intervals applied to data from a T_5 distribution scales with the number of data samples. 55

4.3 How the observed reliability (left plot) and percentile discrepancy (right plot) of the 95% confidence Bayesian tolerance intervals applied to data from a T_5 distribution scales with the number of data samples. 56

4.4 How the observed reliability (left plot) and percentile discrepancy (right plot) of the 99.9 coverage Bayesian tolerance intervals applied to data from a T_5 distribution scales with the number of data samples. 57

4.5 How the observed reliability (left plot) and percentile discrepancy (right plot) of the 95% confidence Bayesian tolerance intervals applied to data from a T_5 distribution scales with the number of data samples. 58

6.1 Flow chart to decide whether and how to present statistical methods for decision-making. 62

6.2 Coverages from the simulation study plotted versus percentile 67

List of Tables

6.1 Percentile estimates using Bayesian and frequentist inference. The true percentile is compared to the median across 500 simulations. 68

Contents

Chapter 1

Introduction

A common objective in reliability engineering is demonstrating with high confidence that a component can meet and has margin to a performance requirement. When component performance measures are continuous, such reliability statements are often made by characterizing the tails of the performance measure distribution using computational simulations or through physical experimentation. When experimental data are collected, the tails of the performance distribution are typically estimated using statistical methods; and margin to the requirement is characterized using probabilistic statements about the estimated tails. Tail characterization plays a key role in fields like probabilistic risk assessment (Atwood et al., 2003; Haimes and Lambert, 1999; Mosleh, 1986) and system reliability and safety (Wallstrom, 2011), with applications in high-consequence settings such as nuclear power plants and nuclear weapons. In such analyses, high system reliability at the top (system) level is achieved by specifying even higher reliability for sub-system level components; however, 'reliability is most sensitive to the tails of these [sub-system] distributions' (Wallstrom, 2011). With high reliability requirements, the sample size is often too small to make nonparametric inferences about extreme tails of a distribution; in this case, parametric distributional assumptions are often used to extrapolate about the extreme tail behavior of the underlying probability distribution (Scholz, 2005).

Quantification of margins and uncertainties (QMU) is a process for quantifying nuclear weapon performance margin and corresponding uncertainty to support risk-informed decision-making about the stockpile (Pilch et al., 2006). In recent years, QMU methods at Sandia National Laboratories (SNL) have relied on characterization of tail behavior using statistical tolerance intervals to inform whether a component has sufficient margin to a requirement. The goal of this report is to characterize risk associated with tail estimation in small to moderate samples.

1.1 Background on QMU for PhysSim at SNL

QMU methods differ across the nulcear weapons laboratories (Eardley, 2005); specifically, QMU at Los Alamos and Livermore National Laboratories pertains primarily to computational simulation due to the data-poor nature of their applications, while QMU at Sandia National Laboratories (SNL) uses more experimental data (called physical simulation or PhysSim data) due to the data-rich nature of Sandia applications (National Research Council, 2008). While the original framework for QMU at SNL was broad conceptual framework (Pilch et al., 2006; Helton, 2009), efforts to move QMU toward experimental data applications (Diegert et al., 2007) have resulted in rather prescriptive, algorithmic processes for PhysSim QMU at SNL (Newcomer et al., 2012). In recent years, the tolerance interval methodology was introduced as the specific methodology to quantify margin in important performance parameters.

PhysSim QMU guidance

The SNL QMU Handbook (Newcomer, 2012) outlines a specific process for implementing PhysSim QMU. The first step is identifying parameters, data, and requirements to use in a QMU; specifically, important performance parameters are identified, their requirements are determined, and then relevant data on the performance parameters are collected and aggregated.

Given a performance measure, requirement, and data, steps outlined in the QMU handbook (Newcomer et al., 2012) for quantifying margin and uncertainty are as follows:

1. Conduct an engineering analysis to assess the data quality.

2. Select a probability distribution for the data.

   The user is advised to use probability plots and distributional hypothesis tests (Anderson-Darling) to select a probability distribution for the data. The user is subsequently warned that, "Not all data are Normally distributed, nor should they be expected to be" (Newcomer et al., 2012).

3. Estimate a tolerance bound and tolerance ratio for the data.

   The QMU handbook recommends using statistical tolerance bounds to estimate margin, defined as the distance between the estimated percentile and CD requirements, and uncertainty, defined as the distance between the estimated percentile and the tolerance bound (Figure 1.1).

   QMU analyses are then directed at answering the question: "Are we XX% certain that atleast YY% of the unit population will yield a response greater than the threshold T?" (Newcomer et al., 2012).

   The recommended 'figure of merit' for decision making is the tolerance ratio, defined as the margin divided by the uncertainty (M/U). Decisions are then based on this M/U metric, noting that, "Comparing the new tolerance ratio against the critical value of 1 is the only decision criteria needed" (Newcomer et al., 2012).

Example implementation of QMU recommendations

To illustrate the PhysSim QMU process, we proceed through the steps using a hypothetical launch safety device as an example. While the example is hypothetical, it is based on our experiences working with component teams to conduct QMU analyses.

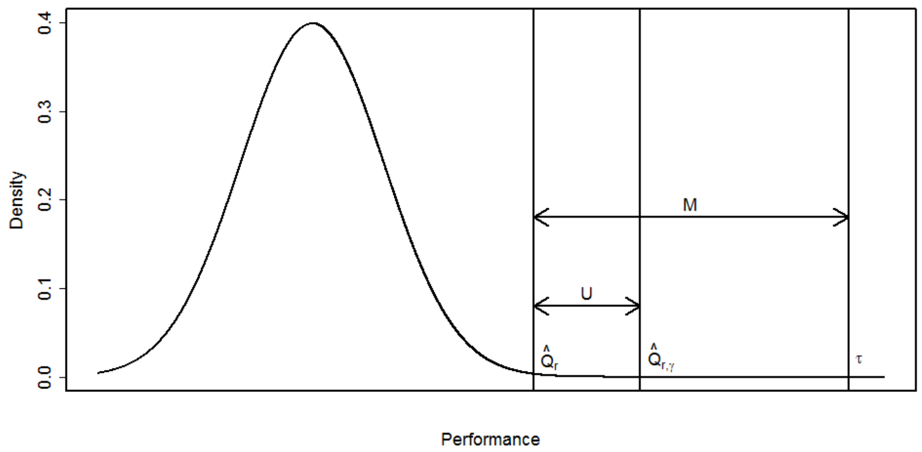

Figure 1.1: Definition of margin and uncertainty in QMU PhysSim process.

Suppose we have 10 closing time measurements (performance parameter) from a launch safety device. For the QMU, we aim to estimate the 99.9^{th} percentile of the closing time distribution with a 95% upper confidence bound to show that we can meet a 34s closing time requirement.

1. **Engineering analysis.** We conduct an engineering analysis on the data based on the handbook recommendations. The data are plotted in Figure 1.2. From an engineering analysis, we conclude that: the data are of sufficient quality but the sample size is small (10 units).

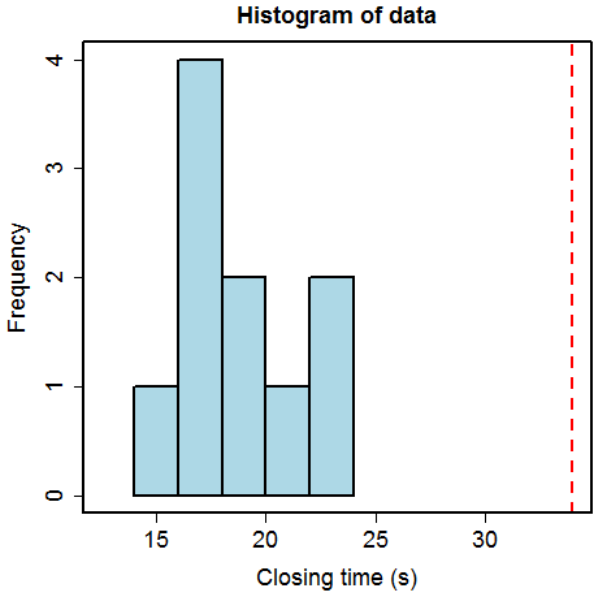

Figure 1.2: Histogram for the $n = 10$ closing time measurements, with the performance requirement in red.

2. **Distributional fit.** We begin with a normal distribution as the candidate probability distribution. We construct a probability plot to graphically assess the normality assumption. If the

data are normally distributed, the theoretical and sample quantiles should fall on a straight line.

Next, we conduct an Anderson-Darling test to test whether there is evidence that the data are not normally distributed. We calculate the p-value to quantify how extreme our observed data are if the data are truly normally distributed; small p-values would suggest a deviation from normality.

The probability plot does not show any obvious deviations from normality (Figure 1.3); the Anderson-Darling p-value is 0.38. From the handbook, "as the p-value is greater than 0.05, we would say that there is not significant evidence at the 0.05 level that the data do not come from a Normal distribution" (Newcomer et al., 2012). Hence, we assume normality and carry on.

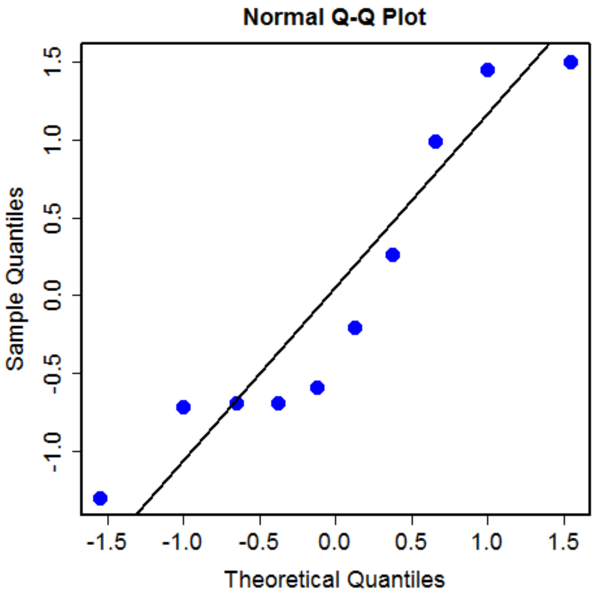

Figure 1.3: Probability plot for $n = 10$ closing time measurements; Anderson-Darling $p = .38$.

3. **Tolerance bound calculation.** Next, we calculated a tolerance bound using the normal distribution to assess "Are we 95% certain that at-least 99.9% of the unit population will yield a response greater than the threshold 34s?"

   From Figure 1.4, we see that the estimated 99.9^{th} percentile is 27s, with 95% tolerance bound 33s. The tolerance bound is less than the CD requirement, and the tolerance ratio is > 1, so we assume margin is sufficient.

Comparison to truth. In this hypothetical example, we generated the data and thus know the correct percentile value. In Figure 1.5, the true data generating model is plotted, illustrating that margin is actually insufficient in this example. The QMU procedure failed to detect and account for model misspecification, leading to anti-conservative and incorrect inferences.

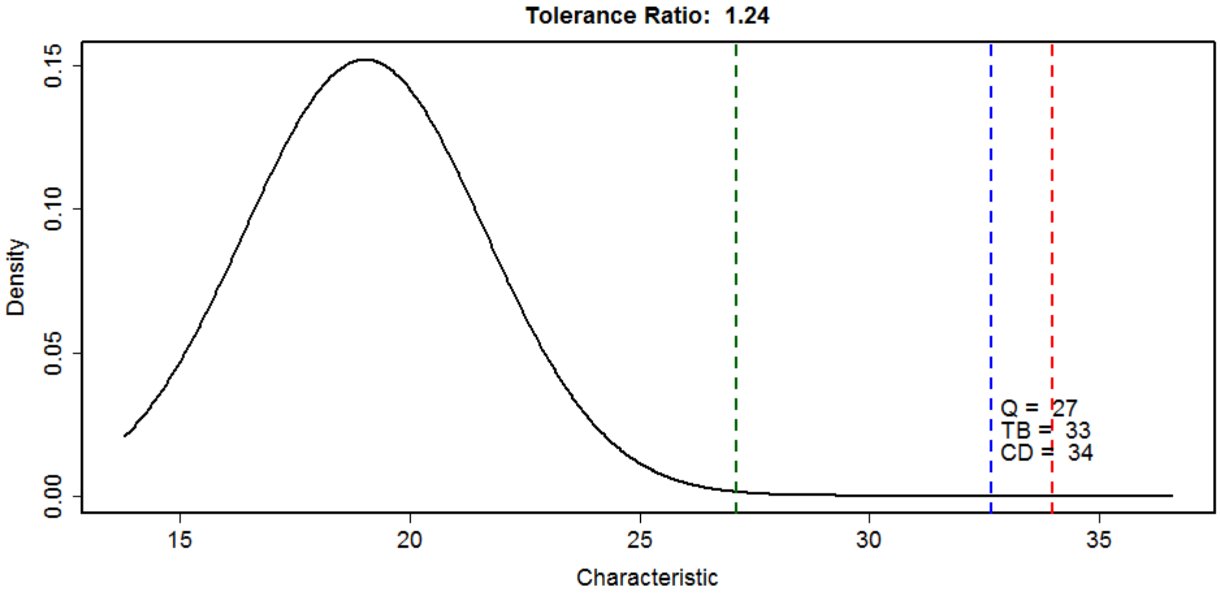

Figure 1.4: Fitted distribution, 99.9^{th} percentile estimate (Q), 95% tolerance bound (TB), and performance requirement (PR) for the $n = 10$ observations. The distance between the TB and requirement (CD) suggests we do not have evidence of a margin insufficiency.

While this example may be hypothetical, such analysis mistakes occur in practice. Recent QMU analyses include statements such as, "There is 95% confidence that 99% of the population will pass" when $1/40$ failed; and "99.999% of the population lies above the [requirement] with 95% confidence" when $n = 150$.

1.2 Perceptions of statistics in PhysSim QMU

In the above example, the PhysSim QMU procedure produced incorrect inferences in the example above for two key reasons:

1. The distributional fit techniques were ineffective at detecting model misspecification.

2. The tolerance interval estimate involved extrapolation far outside the range of the collected data, but statements about margin sufficiency did not address model form uncertainty.

The analysis was **model-driven, rather than data-driven**, and we clearly cannot validate models in the tails without data in the tails. In the sections below, we describe common misconceptions associated with distributional fit techniques and tail extrapolation that lead to such failures of PhysSim QMU in practice.

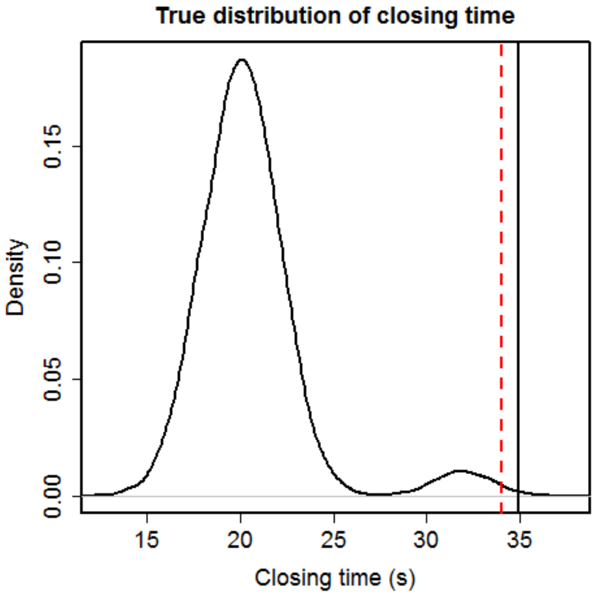

Figure 1.5: True data generating model with vertical line at the true 99.9^{th} percentile (black) and the CD requirement (red). The true percentile exceeds the requirement, and margin is insufficient.

Misconceptions about distributional fit techniques

The problems associated with using probability plots and goodness of fit tests for model validation are well-understood and well-documented. Interpretation of probability plots can be subjective (Loy et al., 2015); is sensitive to outliers (D'Agnostino and Stephens, 1986); and, most importantly for tolerance intervals, highlights the center of the distribution rather than the tails (Scholz, 2005). Distributional goodness of fit tests are even more problematic. D'Agnostino and Stephens (1986) described goodness-of-fit tests as differing from most hypothesis tests in that "it is usually hoped to accept the null hypothesis and proceed with other analyses as if it were true." In 1935, Fisher wrote, "For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning. Yet it does so only too frequently" (Fisher, 1935). In application, goodness of fit tests regularly fall prey to this logical fallacy, which can have severe implications for reliability estimation using parametric statistics. Accepting a parametric model just because the model cannot be disproved using data "can lead to disastrous results" when making inferences about tail behavior (Hughey, 1991).

However, this logical fallacy is sometimes overlooked in QMU applications. Examples include:

- "A statistician can conduct a statistical test to determine whether the observed skew in a random sample is too large to be due to chance... A statistician can also test for kurtosis... In the absence of outliers, bimodality, skewness, and kurtosis, a statistician can determine whether the measurements, or transform of the measurements, are adequately described by a Gaussian distribution" (Diegert et al., 2007).

- “For a critical value α, a p-value greater than α suggests that the data follow that distribution” (Newcomer et al., 2012).

Misconceptions about tail extrapolation

In the example, we used a sample of size 10 to predict the point at which 1/10,000 units will fail (on average). Predicting the behavior of distribution tails can be referred to as ‘tail extrapolation,’ because predictions are not data-driven but are solely reliant on the underlying model. Some dangers of tail extrapolation for tolerance interval estimation are outlined in the quotes below:

- **Probabilistic risk assessment handbook:** “The analyst should not make assertions that are highly dependent on the form of the distribution. For example, a sample of 10 observations may be consistent with many possible distributions. An estimate of the 99.9^{th} percentile of the distribution would be a large extrapolation from the actual data, highly dependent on the assumed form of the distribution. A confidence interval on this percentile would be even worse, because it would give an appearance of quantified precision, when in reality the distribution could have practically any form out in the tail” (Atwood et al., 2003).
- **Los Alamos ONE vs. 1.0:** “[Obtaining] a numerical estimate of reliability based on knowledge of full probability distributions in conjunction with QMU would place great demands on our ability to characterize uncertainties. In view of this, it is inevitable that there would be pressure to adopt ‘short cuts’ by simply assuming the forms of PDFs or using PDFs that are not based on some but inadequate supporting data. The response to such pressure would make or break nuclear certification. No analysis that is based on speculation or that neglects significant possibilities can lead to genuine confidence, but instead will frequently lead to over-confidence or under-confidence, both of which carry severe costs” (Sharp et al., 2003).
- “Estimating tail parameters is analogous to estimating parameters ‘exterior to the data’ ... Many times estimates are made by assuming the data is sampled independently from a parametric family. This can lead to disastrous results” (Hughey, 1991).
- “The tolerance interval... is not distributionally robust to even small deviations from normality” (Fernholz and Gillespie, 2001).
- “Ultimately, the extrapolation step is one of good faith and is not statistical in nature” (Scholz, 2005).

Hahn and Meeker (1982) and Thomas (2015) also discuss the dangers of tail extrapolation for predicting time to failure in industrial applications.

Engineers’ perspectives of tail extrapolation in QMU

We conducted interviews with Sandians to assess perceptions of tail extrapolation in QMU applications and how they use statistical model selection tools in QMU. Seven engineers at SNL in Albuquerque, New Mexico were recruited to be interviewed for the study. Participants were recruited by emailing groups who frequently use QMU methodologies. Eligibility criteria were participants at least 18 years old and employed by SNL for at least one year. All research was approved by the SNL Human Subjects Board.

Methods. Participant interviews lasted between 20-30 minutes; interviews were audio-recorded and transcribed. In the interview, participants were first asked about their experiences with quantifications of margins and uncertainties (QMU) for experimental data in their work. Then, participants were asked about how they use probability plots and goodness of fit tests to evaluate distributional form. Lastly, participants were asked about their perceptions of extrapolation from a statistical model to estimate extreme percentiles (as is common in QMU), and whether goodness of fit tests and probability plots inform whether the extrapolation is reasonable. Using the study transcripts, we identified a set of themes and supporting evidence for the themes.

Results A broad range of QMU users were interviewed. Participants were involved in different areas of the nuclear weapon mission space, including surveillance, production, and development systems. Further, participants had a broad range of technical background in the statistical methods being asked about; all users were familiar with the statistical methods, because they are tied to the QMU methodology. While the users varied in their understanding of statistics, all participants were in some degree responsible for interpreting and reporting out findings from QMU studies that use these methods.

We identified three primary themes from the focus group. Specifically, participant responses generally fell under the following themes:

Theme 1. *Using QMU as an exercise to think about margin generates valuable information.* QMU was deemed useful for: helping think through the quality of available data; understanding data in manufacturing and testing; evaluating performance in surveillance, including Annual Assessment Reviews; and communicating with customers about observed performance. Going through the QMU process was also deemed useful for understanding shifts or outliers in data, identifying subpopulations, and, ultimately, understanding variability to better understand the probability of exceeding requirements. The process was deemed easy to use to screen out performance parameters with high margin.

Engineering (data) reviews were deemed a useful part of the QMU exercise. Participants noted the need to understand what data are available, what the requirements are, and whether QMU is realistic and feasible.

Theme 2. *Extrapolation in QMU is a source of risk.* In general, participants understood that extrapolation from a parametric statistical model was a risky endeavor. Example participant responses include: "we're making predictions about events we haven't seen yet" and "we really havent collected any data points there yet." Participants noted: there is rarely enough data to prove or disprove modeling assumptions, outliers are always a concern, and that the risk associated with assumptions must be articulated. However, one participant highlighted that briefings to customers often do not include risks associated with assumptions. The perceived magnitude of risk seemed to vary between participants. Example participant responses that down-played the risk include:

“data in the tails tend to be similar” and statistical confidence is “necessary and sufficient” to characterize uncertainty following a thorough engineering analysis. Further, many participants had a good understanding that the level of rigor needed in an analysis depends on how much margin exists. Specifically, high margin situations need less rigorous methods. One participant noted that more qualitative tools may provide added-value for mitigating the risk of extrapolation.

Theme 3. *Statistical tools for evaluating parametric distribution fit (Anderson-Darling goodness of fit tests and probability plots) are sufficient to identify a distributional model to use for QMU extrapolation.* Participants seemed to have good faith in goodness of fit tests and probability plots for selecting a distributional form. Example participant responses include: these methods are the ‘gold standard’; ‘Minitab has functions that will identify a distribution for you’; and statisticians can tell what distributional form data have (paraphrased). There was often a lack of skepticism concerning statistical methods. The statistical tools were not related to the degree of extrapolation being done (i.e. the percentile and desired confidence level); for instance, no one was concerned about extrapolating a 99.5th percentile with 95% confidence using a sample size of $n = 65$.

From these qualitative interviews, we learned that more intuitive statistical tools are needed to communicate the risk of extrapolation when modeling experimental data.

1.3 Objectives of report

In this report, we highlight the risk associated with the current PhysSim QMU procedure, discussing both the likelihood and consequences of model misspecification. We then propose a path forward for PhysSim QMU based on information integration. Specifically, we address the following three objectives to highlight risk associated with PhysSim QMU processes:

1. **Quantify model form risk:** Quantify the consequences of model misspecification using a scaling study (consequences).
2. **Define model validation metrics:** Develop improved model validation metrics (likelihood).
3. **Propose information integration models:** Explore potential solutions for relaxing stringent model form assumptions through integrating multiple data sources (risk mitigation).

Because these three objectives represent distinct aspects of the project, we have structured the document to group the objectives together. Specifically, we present methods, results, and discussion for each of these objectives. In Section 2, we present the scaling study to quantify model form risk; in Section 3, we present the model validation metrics to evaluate the likelihood of model misspecification in the tails; and in Section 4, we outline a potential data integration framework for risk mitigation.

Chapter 2

Quantify model form risk

When estimating uncertainty in percentile estimates (tolerance bounds), there are two sources of uncertainty to consider in analysis: uncertainty associated with model misspecification and sampling uncertainty assuming the model is correct. Sampling uncertainty is conditional on the model, but can be quantified for models that make various different parametric assumptions; naturally, sampling uncertainty increases as models become less parametric (i.e., less restrictive), because fewer assumptions can be leveraged to characterize the underlying population. In statistics, this phenomenon is referred to as the bias-variance trade-off.

Sampling uncertainty is quantified when estimating statistical tolerance intervals, but model uncertainty is frequently ignored, particularly in the SNL PhysSim tolerance interval methodology (Newcomer et al., 2012). When highly parametric models are used (e.g. normal, Weibull, lognormal), understanding the strength of the modeling assumptions can be achieved by examining the performance of the parametric models when the model is wrong. In this section, we quantify the impact of both sampling and model uncertainty on tolerance interval estimates.

2.1 Methods

To highlight the impact of model and sampling uncertainty on tolerance interval estimates, we propose examining:

1. When the parametric model is misspecified, how does the true confidence compare to the nominal confidence level?
2. When the parametric model is misspecified, how does the tolerance bound compare to the true percentile?

The difference between the confidence level prescribed by a tolerance interval and that observed is termed empirical reliability and will provide evidence for the first question. The mean difference between the true percentile and the tolerance bound estimate is termed the percentile discrepancy and this will provide evidence for the second question. Using these two metrics, we quantify model and sampling uncertainty in tolerance bound estimates across different sample sizes, target percentiles, confidence levels, and underlying models.

To understand empirical reliability, note that tolerance intervals have a confidence level associated with them, say 95%. This confidence level can be interpreted as the frequency with which the interval contains the true parameter *assuming the assumptions of the statistical method are met*. A common statistical assumption is that the selected statistical model is correct; if the model is wrong, we are no longer guaranteed that the 95% confidence interval will contain the true parameter 95% of the time. Tolerance intervals are particularly sensitive to model form; for instance, assuming a model that is too light-tailed will produce under-conservative results (true confidence less than 95%).

For this study, we generate data assuming that the true underlying model is known. In practice, with experimental data, we do not know the true model that generated the data resulting in model uncertainty; this study is intended to show the sensitivity of tolerance bounds to model form assumptions. To accomplish this we take random samples from a known distribution and estimate a tolerance interval assuming a mis-specified model. We then compare the estimated tolerance interval to the true percentile; and compare the nominal confidence level for a large ensemble of random samples to the estimated coverage of the true percentile under the correct model.

Study design

We designed a study to explore the metrics. We **assume a normal model** when calculating the tolerance interval, but vary the true underlying model, considering the correct normal model as well as two incorrect alternative models: a t-distribution with 5 degrees of freedom (T_5) and a mixture of 2 normal distributions. The normal distribution reflects a 'light-tailed' distribution, t-distribution is intended to reflect a heavy-tailed distribution, and the normal-mixture a multi-modal distribution. The normal-mixture examples was selected to reflect a real application to computational simulation data from a launch safety device on a nuclear weapon, but we omit details of the specific application herein.

We consider the following design variables throughout the simulation study:

- different sample sizes: $n = 4, 6, 8, ...40$.
- 4 different quantiles $p = 0.90, 0.99, 0.999, 0.9999$.
- 4 confidence levels $\gamma = 0.5, 0.9, 0.95, 0.99$.

We generate one dataset (a large ensemble of random samples) for each setting (distribution and sample size) and calculate the true confidence interval coverage and the difference between the average tolerance intervals and true percentiles.

2.2 Results

In this section, we explore the risk of model form assumptions on two relevant distributions that highlight practical issues with model form assumptions in sparse data situations. Before highlighting the risks in making model form assumptions, the appropriate application of the normality assumption to normally distributed data is shown to set a baseline for the performance metrics. Although normality is not assumed for all analysis, these examples are meant as a general illustration of the risks associated with making model form assumptions.

2.2.1 Correct model form assumption: Normal distribution

To illustrate our approach to characterizing the performance of model form specific percentile estimation, tolerance intervals (TIs) based on the normal distribution are applied to normally distributed data $\mathscr{N}(0, 1)$. For this illustration 30,000 combinations of samples from the normal distribution are considered for each number of sample points considered. Each combination of samples is tested with the Anderson-Darling (AD) test at a 0.05 p-value to determine if the hypothesis that the data comes from a normal distribution can be rejected. Figure 2.1 shows the type I error rate of the AD test applied to datasets from a normal distribution. The type I error rate is defined as the probability of rejecting the hypothesis that the data comes from a normal distribution when the normal model is correct. The type I error rate appears to operate around at a 0.05 rejection rate, which is expected due to the data being truly normally distributed and the hypothesis test rejecting at a 0.05 level. If the hypothesis that the data is normally distributed is not rejected, a one sided tolerance intervals is fit to the data and compared to the true percentiles of the standard normal distribution.

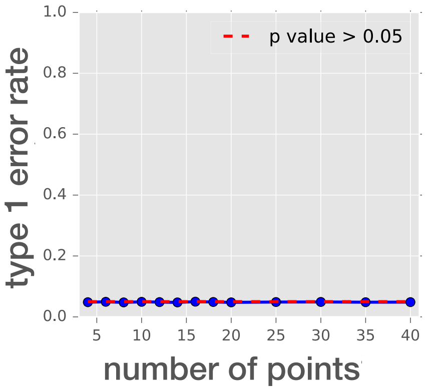

Figure 2.1: Type 1 error rate scaling of Anderson-Darling test for normality applied to data from standard normal distribution at a 0.05 p-value

Performance metric results for the normal distribution based tolerance intervals (TI) applied to normally distributed data are shown in Figure 2.2 and 2.3. The empirical reliability is the difference between the prescribed confidence and the observed. Figure 2.2 shows how the performance metrics scale with the number of sample points for 99.9% coverage tolerance intervals with different confidence levels. Ideal empirical reliability occurs when the observed confidence equals the specified confidence level. It appears that the observed results are approximately ideal given the limited number of repetitions. Discrepancy in the percentile estimates is expect to approach zero as the number of samples increases. A trend in the percentile discrepancy towards zero is evident and the magnitude of the discrepancy for small sample sizes scales with the magnitude of the estimated percentile. Figure 2.3 then similarly shows the scaling of the performance metrics for 95% confidence tolerance intervals with different coverage levels. Again, the ideal reliability would be achieved if the observed and prescribed confidence levels matched, which appears to occur for the normally distributed data. The percentile discrepancy appears to scale in a manner consistent with Figure 2.2, and the percentile discrepancy has a positive trend with the magnitude of the confidence level. Thus, when applying these performance metrics, it is hoped that the prescribed confidence level is matched, the percentile discrepancy should asymptote to zero as the number of samples increases, and the magnitude of the percentile discrepancy should increase with higher estimated percentiles and confidence levels (but with a stronger trend with respect to percentiles).

Summarizing these results, when the normal model is correct, normal tolerance intervals work well for bounding distribution percentiles, as expected.

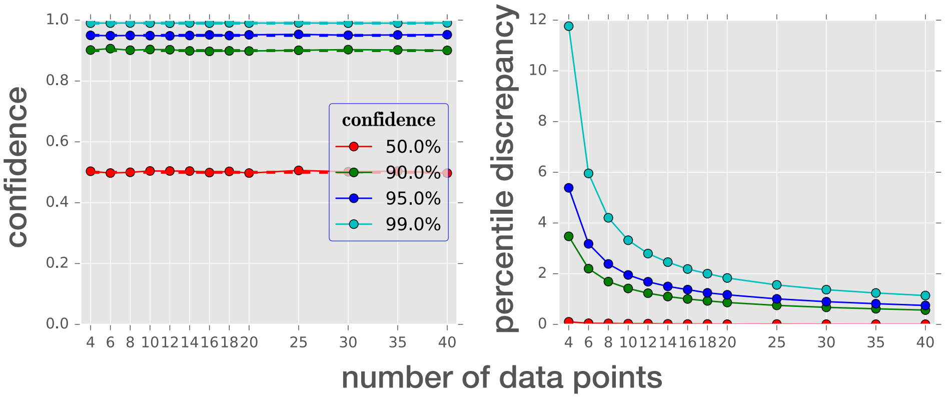

Figure 2.2: Scaling of mean performance metrics across sample size for 99.9% coverage tolerance intervals (assuming normality) for data from a standard normal distribution. Prescribed confidence levels of the tolerance intervals (dashed lines) are compared with the observed confidence (empirical reliability) in the left plot, and the percentile discrepancy is shown in the right plot. All values are statistical means over a large population of random samples.

2.2.2 Incorrect model form assumption: Tail sub-population

The first distribution used to illustrate issues with tolerance interval robustness to model form uncertainty appears roughly normal, but contains a subpopulation in the right tail, as shown in Figure 2.4. As an illustration of the impact of the tail subpopulation, the 99.9 percentile of the distribution is compared with that of a normal distribution fitted to the mean and standard deviation of the distribution. For each desired sample size, 20,000 combinations of samples are taken from the distribution and normal distribution based tolerance intervals are fitted to each combination.

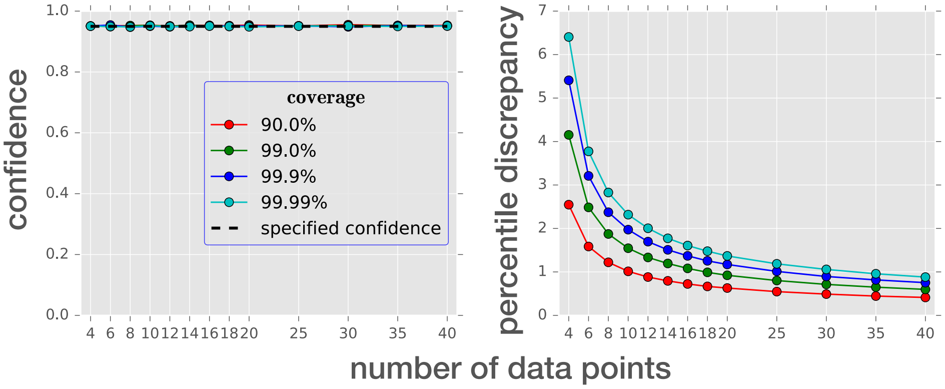

Figure 2.3: Scaling of mean performance metrics across sample size for 95% confidence tolerance intervals (assuming normality) for data from a standard normal distribution. Prescribed confidence levels of the tolerance intervals are compared with the observed confidence (empirical reliability) in the left plot, and the percentile discrepancy is shown in the right plot. All values are statistical means over a large population of random samples.

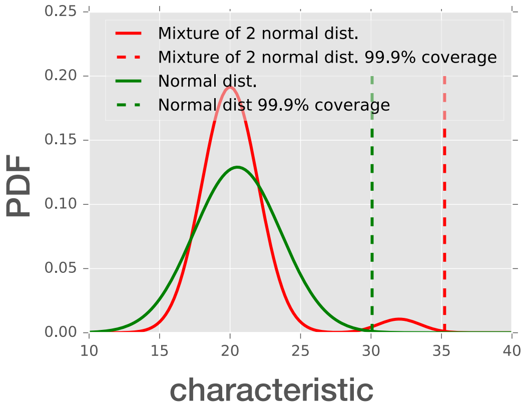

Figure 2.4: Distribution with right tail subpopulation (red solid line) and 99.9 percentile (red dashed lines) compared with 99.9 percentile based on equivalent normal distribution.

The sample datasets are filtered using the AD normality test to ensure only datasets that would not be rejected are used to determine the tolerance intervals' performance. Figure 2.5 illustrates how the statistical power of the AD test with a 0.05 p-value scales with the number of sample data points. Statistical power is defined as the probability of rejecting the hypothesis that the data comes from a normal distribution when the true distribution is not normal and, in this case, is the normal mixture distribution. With small sample sizes the AD test rejects few of the incorrect hypotheses for this example, but the power grows with the number of sample points, as expected. With datasets of 18 samples, it is expected that around 50% of the time the dataset will be determine to not be from a normal distribution.

To get a more complete picture of how the performance of normal tolerance intervals scales for this underlying distribution, Figure 2.6 and 2.7 demonstrate the performance scaling across a range of confidence levels, distributional coverages, and number of sample points. The empirical

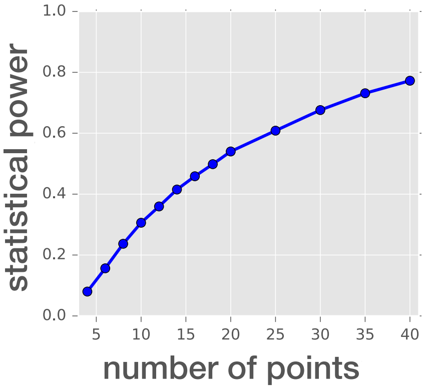

Figure 2.5: How the rejection rate of the AD test at a 0.05 p-value, scales with number of sample points for random combinations of datasets taken from the non-normal population shown in Figure 2.4.

reliability of all confidence levels for 99.9% coverage TIs performed best with sparse data, but still did not achieve 'nominal' coverage. The confidence performance quickly reduces as additional data points are added until it starts to asymptote to zero confidence when the sample size is around ten to twelve points. With ten to twelve sample points, the statistical power plot indicates that around 70% of datasets are not rejected by the AD normality test. In sparse data situations, the 99.9 percentile estimates are typically overly conservative on average, except for the estimates based on 50% confidence. As the number of sample points grows, the percentile discrepancy quickly becomes negative (anti-conservative). The 99.9/50 TIs are underestimated for all sample sizes.

Within the coverage scaling (Figure 2.7), the impact of the main distribution on the 90% coverage TI is evident. Because the subpopulation made up less than 10% of the total PDF, the normal distribution approximation TI performed much better for the 90/95 TI than for the higher percentile based TIs that are located on the tail of the subpopulation. This impact is notable in both the empirical reliability and the percentile discrepancy. The normality assumption is nonconservative for all percentiles considered, but performs better for less extreme percentiles located not on the subpopulation.

Summarizing the results, if a subpopulation exists in the data and enough samples were not taken to adequately reject the normal model, normal model inferences will be anti-conservative.

2.2.3 Incorrect model form assumption: T_5 distribution

The second test distribution used to illustrate issues with tolerance interval robustness to model form uncertainty was a T distribution with five degrees of freedom (T_5) whose position and scale (similar to normal distribution mean and standard deviation) are uncertain. The T_5 distribution has a shape close to a normal distribution, but has thicker tails. Figure 2.8 compares the standard T_5 distribution with a standard normal distribution. Looking at the 99.9 percentiles illustrates the

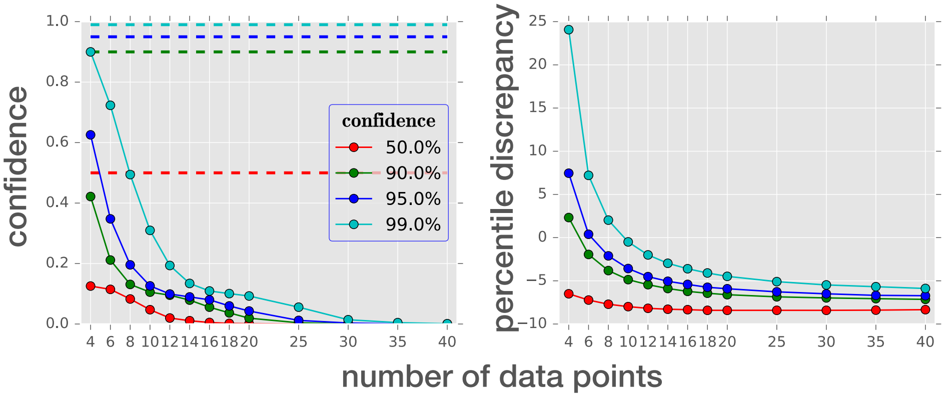

Figure 2.6: Observed reliability (left plot) and percentile discrepancy (right plot) of the normal 99.9% coverage tolerance intervals applied to data from the distribution with a subpopulation on the right tail (Figure 2.4). Trends depending on different prescribed confidence levels (50%, 90%, 95%, and 99%) are shown as solid lines with dots and corresponding prescribed confidence levels are shown as dashed lines. Values shown are average values from 30,000 random combinations, but only combinations not rejected by AD test at a 0.05 p-value are considered.

greater weight in the tails of the T distribution.

The AD test with a 0.05 p-value was again used to screen 30,000 combinations of data points sampled from the T_5 distribution to get the empirical reliability and percentile discrepancy estimates. The power of the AD normality hypothesis test is shown in Figure 2.9. Due to the similarity of the T_5 and normal distribution (aside from the distribution tails), the power of the AD test is significantly reduced for small sample sizes. With a sample size of 30, the hypothesis that data from a T_5 distribution was actually from a normal distribution was only rejected for ≈20% of datasets.

Figure 2.10 and Figure 2.11 show how the performance of normal distribution based TIs to estimate percentiles of the T_5 distribution scales across number of samples with different coverage and confidence levels. Just as was noted for the distribution with a tail subpopulation, the normal distribution based TIs have their best coverage performance with sparse data and performance then trends down as additional data points are included. The rate of decrease in the TI confidence performance is slower than was observed for the distribution with a tail subpopulation. The 50% confidence 99.9% coverage TI again is anti-conservative on average for all sample sizes and all TI confidence levels considered trend to a negative percentile discrepancy.

The scaling of the empirical reliability and percentile discrepancy with coverage level also scales in a similar fashion to the distribution with a subpopulation, but with more gradual slopes. The exception to this is the 90/95 TI which performs at the specified confidence level for the sample sizes investigated. This is likely due to the T_5 and standard normal distributions being similar for percentiles not far into the tails.

Summarizing the results, it is often not possible to detect distribution tails that are heavier than the normal distribution using statistical goodness of fit tests, but differences in distribution tails have important impacts when predicting tail behavior.

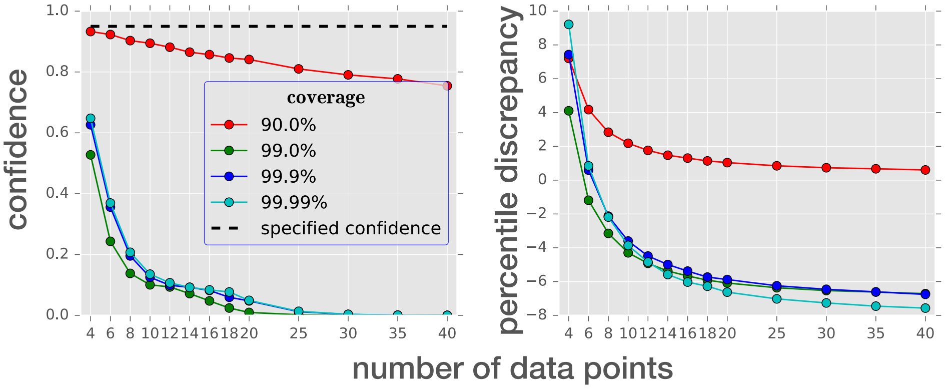

Figure 2.7: Observed reliability (left plot) and percentile discrepancy (right plot) of the normal 95% confidence tolerance intervals applied to data from the distribution with a subpopulation on the right tail (Figure 2.4). Trends depending on different prescribed coverage levels (90%, 99%, 99.9%, and 99.99%) are shown as solid lines with dots and corresponding prescribed confidence levels are shown as dashed lines. Values shown are average values from 30,000 random combinations, but only combinations not rejected by AD test at a 0.05 p-value are considered.

2.3 Discussion

Although it is known that tail extrapolations are sensitive to model form assumption, alternative procedures are not well developed for sparse data situations. In engineering applications where QMU type questions are being asked, analysts often rely on the most accessible tool, model form assumptions. Results in section 2.2 were generated to help quantify the potential risks being taken when making model form assumptions to perform QMU. Assuming normality is a common engineering assumption, so this assumption was applied to realistic distributions that deviate from normality, which engineers are likely to encounter in real applications. Grounding the two performance measures, observed reliability and percentile discrepancy, on the performance of the normality model form assumption for normal data provides both insurance of the algorithms implementation and a baseline for the optimal values for the metrics.

Applying AD goodness of fit tests at the significance level suggested within Newcomer et al. (2012) to the test problems illustrated issues with this method in sparse data situations. The performance of the normal tolerance intervals applied to both test problems was poor, and a tradeoff between confidence and bias in the percentile estimates was apparent. Incorrect model form assumptions act more conservatively in their confidence of truly bounding the desired percentiles when data is most sparse, but this is due to the large overestimates in the percentiles, as shown in the percentile discrepancies. Specifically, with sparse samples, variance estimation is challenging and, thus, uncertainty in the variance under the normal model dominates uncertainty in this case. As data becomes less sparse, over-prediction of percentiles appears to asymptote quickly as more data is added. The negative impact of the model form assumptions is evident in the poor confidence in percentile estimates due to negative percentile discrepancies as additional data is added. Even with sparse samples, the empirical reliability is far from the nominal statistical confidence (as expected). Using as much data as possible allows the goodness-of-fit tests the best chance of correctly

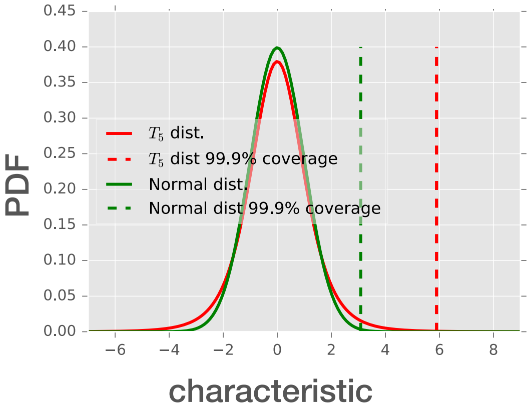

Figure 2.8: T_5 distribution (red line) compared to normal distribution equivalent (green line). 99.9 percentiles of the distributions are shown as dashed lines.

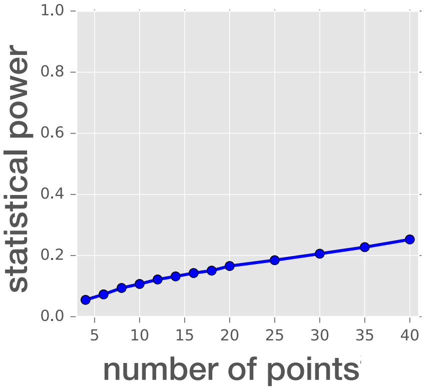

Figure 2.9: How the rejection rate of the AD test at a 0.05 p-value, scales with number of sample points for random combinations of samples taken from non-normal population shown in Figure 2.8.

identifying a poor model form assumption, does not significantly increase the expected discrepancy from what would be experienced for most sparse data levels, but does greatly increase the risk associated with the confidence prescribed. However, goodness of fit tests will almost always reject the distributional form assumptions in large samples, given that data never exactly follow prescribed distributions, even though those distributions are often reasonable approximations.

Comparing the performance of the incorrect model form assumption on the two test distributions considered, it appears that the normality assumption performs worst for non symmetric distributions. The decline in the observed reliability for the distribution with a subpopulation in one of its tails was significantly worst than for the systematic T_5 distribution. On the other side of this coin, the normality assumption appears to perform relatively well for more central and symmetric portions of the distributions. This knowledge may prove useful for other engineering activities such as calibration, where greater emphasis is placed on capturing central behaviors such as those studied by Romero et al. (2013).

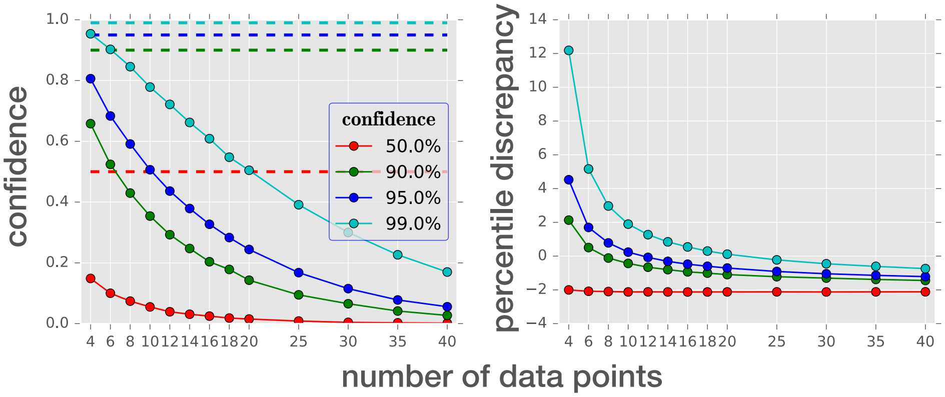

Figure 2.10: Observed reliability (left plot) and percentile discrepancy (right plot) of the normal 99.9% coverage tolerance intervals applied to data from a T_5 distribution. Trends across sample size depending on different prescribed confidence levels (50%, 90%, 95%, and 99%) are shown as solid lines with dots and corresponding prescribed confidence levels are shown as dashed lines. Values shown are average values from a large number of combinations, but only combinations not rejected by AD test at a 0.05 p-value are considered.

The performance plots for the two examples provide analysts with a simple visual assessment of the potential impact of poor model form assumptions commonly employed. For example, if the analyst's true distribution was the distribution with a subpopulation in its tail, around 56% of the time 20 points randomly taken from that distribution would be assumed to be normal and not proven otherwise. The 99.9% coverage/99% confidence tolerance intervals based on those 20 points would then bound the 99.9^{th} percentile around 10% of the time and would have an estimate discrepancy around 4 units under the true percentile value.

Another finding from this study is that 50% confidence should be avoided. In some applications, analysts prefer 50% confidence because fewer samples are required. However, 50% confidence in a percentile is essentially just a point estimate of the percentile and ignores sampling uncertainty. This study highlighted the consequences of ignoring that uncertainty in terms of very poor empirical reliability and large percentile discrepancies. Hence, we recommend that 50% confidence not be used as the confidence level for tolerance intervals.

One means of overcoming the risk associated with model form assumptions is to change the type of QMU questions being asked in sparse data situations. Instead of asking for margin ratio estimates based on extreme percentiles, the amount of data available could dictate the percentiles being estimated. Where perviously the margin ratio based on the 99.99 percentile of a performance characteristic would be requested from a dataset containing 50 points, what if a percentile less than 1/50 was used to report the margin ratio? Segalman et al. (2017) published an approach to QMU where the data was first shifted so that a small percent of the distribution was beyond the requirement, and the necessary shift was referred to as the margin. Probabilities of failure were then based on the margin shifted data, resulting in estimates largely insensitive to model form assumptions. Ultimately, the questions being posed by decision makers would need to accommodate this change in mindset before QMU approaches could be altered.

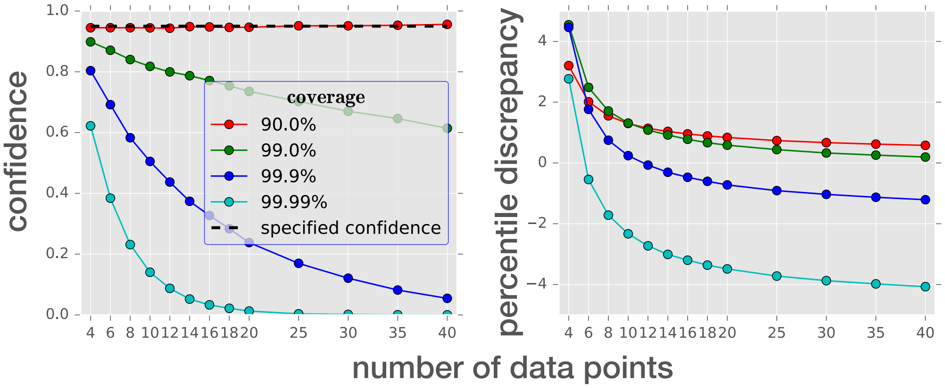

Figure 2.11: Observed reliability (left plot) and percentile discrepancy (right plot) of the normal 95% confidence tolerance intervals applied to data from a T distribution with 5 degrees of freedom. Trends depending on different prescribed coverage levels (90%, 99%, 99.9%, and 99.99%) are shown as solid lines with dots and corresponding prescribed confidence levels are shown as dashed lines. Values shown are average values from a large number of combinations, but only combinations not rejected by AD test at a 0.05 p-value are considered.

Chapter 3

Define model validation metrics

We construct alternative and improved metrics for communicating model form risk when considering whether to estimate distribution tails from statistical models to characterize rare event probabilities. Specifically, we use statistical extreme value methods to communicate risk associated with tail characterization for assessing high reliability requirements. Statistical extreme value theory is concerned with making inferences about extreme observations (Coles et al., 2001) and thus provides a suite of tools relevant for reliability analysis with high requirements. We propose graphical aids and statistical metrics to address the question of when enough data is available to validate a parametric model for extreme percentile estimation. The proposed tools parallel probability plots and goodness of fit tests but directly explicate the risk being absorbed in model-based extrapolation.

To develop an alternative to goodness of fit tests, we develop a validation metric that is directly tied to validation of physics-based computational simulations (AIAA, 1998; Oberkampf and Barone, 2006; Rebba and Mahadevan, 2008). While the precise definition of model validation is debated in computational simulation applications, we will define validation as, “the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model” (Oberkampf and Barone, 2006). In short, validation examines whether a model’s prediction accuracy is good enough for the intended application. Statistical model validation is no different from computational simulation validation; therefore, to validate statistical models for the intended application, i.e. tail characterization, we should provide evidence that we can accurately capture observed tail behavior with the model. We argue that model validation in statistical modeling should not be treated differently from validation in computational simulation modeling and provide metrics aimed at bridging the gap between current treatments of model validation in these fields.

3.1 Methods

3.1.1 Percentile and tolerance interval estimation

To define the tail characterization problem, we consider estimation of an extreme percentile of a distribution; in practice, this percentile of interest may correspond to a reliability or safety requirement. For a component within a system, suppose the component has a requirement that a performance measure must be less than a performance threshold τ with reliability r. That is, the Q_r percentile of the performance metric distribution must be less than τ. (Throughout, we consider an upper performance threshold for sake of simplicity, but all methods apply to both upper and lower thresholds.) To characterize confidence (or, conversely, sampling uncertainty) in percentile estimates, we use statistical tolerance intervals (Krishnamoorthy and Mathew, 2009). We first review percentile and tolerance interval estimation, before describing proposed metrics for tail estimation.

Percentile and tolerance interval estimation

Without loss of generality, we consider percentile estimation for the upper-tail of a probability distribution. Further, we assume all performance measures are independent and identically distributed ($i.i.d.$), a common assumption in practice. Let $X = \{X_1, X_2, ..., X_n\}$ denote an $i.i.d.$ random sample from a population. Given a set of data X, we can estimate the distribution of X, denoted $\mathscr{F}_X$, and then calculate percentiles from this distribution. A percentile of a distribution Q_r is defined as the value of the distribution for which r percent of observations are below. Mathematically, Q_r is defined as $P(X < Q_r) = r$.

In practice, the distribution $\mathscr{F}_X$ and percentiles Q_r are estimated from finite samples and contain sampling uncertainty. We denote estimates of $\mathscr{F}_X$ and Q_r as $\hat{\mathscr{F}}_X$ and $\hat{Q}_r$. Statistical tolerance intervals can be used to quantify sampling uncertainty in a percentile $\hat{Q}_r$. Tolerance intervals can be one- or two-sided; we consider only one-sided tolerance intervals in this paper, often referred to as tolerance bounds, because, in practice, we are typically trying to find a bound on a performance measure. Mathematically, a one-sided $(p, 1-\alpha)$ tolerance bound $\hat{Q}_{p,\alpha}$ is defined as:

$$P_{\mathbf{X}}\left\{P_X\left(X \leq \hat{Q}_{p,\alpha}|\boldsymbol{X}\right) \leq p\right\} = 1-\alpha, \tag{3.1}$$

where p is a percentile, α is the confidence in the estimate of the percentile. Heuristically, a one-sided upper statistical tolerance bound is simply an upper confidence bound on a percentile.

Non-parametric estimation

With sufficient data, assumptions about the underlying model $\mathscr{F}_X$ are not necessary for percentile and tolerance interval estimation. Percentiles are estimated using quantiles of the observed data. Non-parametric one-sided tolerance intervals can be constructed based on exact distributions of order statistics (Scholz, 2005; Krishnamoorthy and Mathew, 2009; Wilks, 1941; Hutson, 1999). To help gauge how much data is sufficient for non-parametric estimation, we use the minimum sample size requirements for non-parametric tolerance interval construction. Specifically, to construct a $(p, 1-\alpha)$ nonparametric interval, n must satisfy

$$n \geq \frac{log(\alpha)}{log(p)} \tag{3.2}$$

where n is the sample size, $p > 0.5$ is the percentile of interest, and $1-\alpha$ is the confidence level (Wilks, 1941). Non-parametric tolerance intervals make no assumptions about the shape of the underlying distribution at the cost of requiring very large sample sizes for construction. For instance, bounding a 99^{th} percentile with 95% confidence requires $n \sim 300$; bounding a 99.9^{th} percentile with 95% confidence requires $n \sim 3,000$.

Parametric estimation

When the constraint in Equation 3.2 is not met, extrapolation based on a statistical model is commonly used to estimate percentiles and tolerance intervals, i.e. $X_i \sim \mathscr{F}_X$, where $\mathscr{F}_X$ is an assumed probability model for X. In engineering applications, common choices for $\mathscr{F}_X$ are the normal, Weibull, and lognormal distributions (Atwood et al., 2003; Krishnamoorthy and Mathew, 2009; Newcomer et al., 2012). Parameters of the distributions are estimated using statistical methods, such as maximum likelihood estimation or method of moments. After fitting the model, percentiles are estimated based on quantiles of the fitted model $\hat{\mathscr{F}}_X$.

Extreme-value methods

Often, it is difficult to find a parametric distribution that fits the entire observed dataset well. To relax the assumptions of parametric approaches, extreme-value methods can be applied, where only the m largest observations from a dataset are used to make inference about the upper tail behavior. Extreme-value methods are commonly used to estimate extreme percentiles and corresponding tolerance intervals (Hughey, 1991; Coles et al., 2001; De Haan and Ferreira, 2007; Gomes and Guillou, 2015). The basic notion behind this approach is that, if we are interested in inference about the tails, then only the tails should be used to make inference. To implement tail-based methods, the data are censored at some threshold u and inferences about the tails of the distribution of X are based on estimating the distribution $Y = X - u | X > u$. Extreme value methods are inherently less parametric and thus more robust than the parametric approaches described above, again at the cost of efficiency. A common model for distribution tails is the generalized Pareto distribution (referred to as the 'peaks over thresholds' method) (Coles et al., 2001). The distribution function for the generalized Pareto distribution (GPD) is:

$$F(y) = 1 - (1 + y\gamma/\sigma)^{-1/\gamma} \tag{3.3}$$

where $\gamma \neq 0$ and $\sigma > 0$ are shape and scale parameters, respectively, and Y has support > 0 if $\gamma > 0$ and $[0, -\sigma/\gamma]$ if $\gamma < 0$. If $\gamma = 0$, $F(y) = 1 - exp(-y/\sigma)$. Other more common distributions, such as the normal, Weibull, or lognormal distributions, can also be used to model distribution tails. There is a clear bias-variance trade-off involved in selecting how much data to use to model the tails, i.e. selection of u; choosing u too large will result in very little data being used for estimation, while choosing u too small will result in data further from the tails influencing the tail fit.

3.1.2 Motivating example: QMU for launch safety device

To illustrate practical challenges associated with tail characterization, we consider an example motivated by quantification of margin and uncertainty (QMU), a framework for evaluating system-level nuclear weapon performance (e.g. reliability and safety) by rolling-up component level margin estimates (Pilch et al., 2011; Newcomer, 2012). Given a reliability requirement r, the estimated r^{th} percentile of the performance distribution ($\hat{Q}_r$) must be sufficiently far from the requirement τ, accounting for uncertainty (Figure 3.1). In QMU, margin is defined as the distance between the percentile estimate $\hat{Q}_r$ and requirement τ; (sampling) uncertainty in the percentile estimate is measured through a tolerance interval $\hat{Q}_{r,\alpha}$. We assume the requirement τ is fixed and known. As the demand for QMU using experimental data increases, a common question is when enough data exists to reliably estimate a percentile Q_r and a tolerance interval $\hat{Q}_{r,\alpha}$. More precisly, when can we feel confident that a statistical model $\mathscr{F}_X$ is valid for estimating a percentile and tolerance interval?

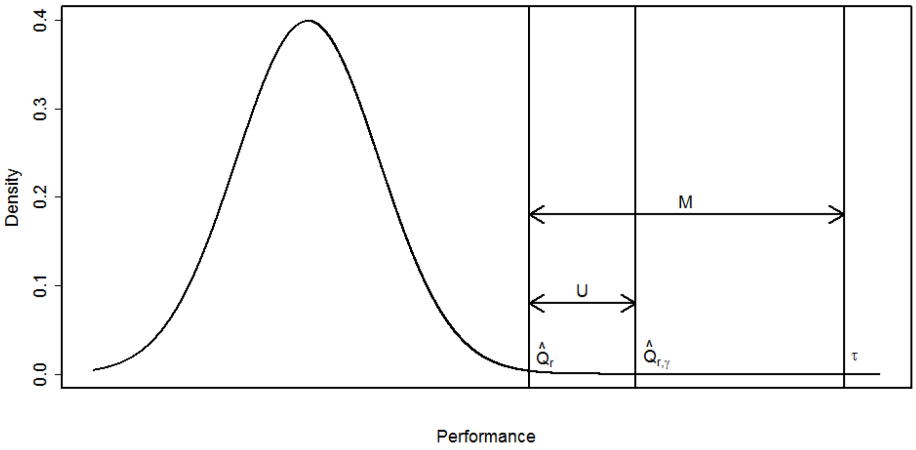

Figure 3.1: Illustration of a percentile estimate $\hat{Q}_r$ and a one-sided 95% confidence bound on the percentile estimate $\hat{Q}_{r,.95}$. The margin M to the requirement τ is the distance between the bound $\hat{Q}_r$ and the requirement threshold τ. Using this margin-based approach, if $\hat{Q}_{r,.95}$ is less that τ, then we have at least 95% confidence that the requirement is met under the assumed statistical model.

As an example, we consider a hypothetical launch safety device on a missile. Suppose the component has a requirement to close within 23.5s of launch with 99.5% reliability; that is, we aim to estimate the 99.5^{th} percentile of the closing time (CT) distribution to find evidence supporting that the component can meet the $\tau = 23.5s$ requirement. The component is tested at two different settings: hot and cold temperature; cold CT is expected to be more variable due to increased friction. From both temperature settings, we have $n = 100$ closing time measurements (Figure 3.2); we refer to these two performance outcomes as hot CT and cold CT. While all data used in this manuscript are simulated, the examples are motivated by real but proprietary examples.

Both distributions look reasonably symmetric and bell-shaped; in such situations, normal distributions are commonly used as a probability model for data. Using the normal distribution to estimate a 95% confidence interval on the 99.5^{th} percentile (99.5/95 tolerance interval), the estimated tolerance interval for hot CT is 22.8s and for cold CT is 23.3s. Therefore, based on the

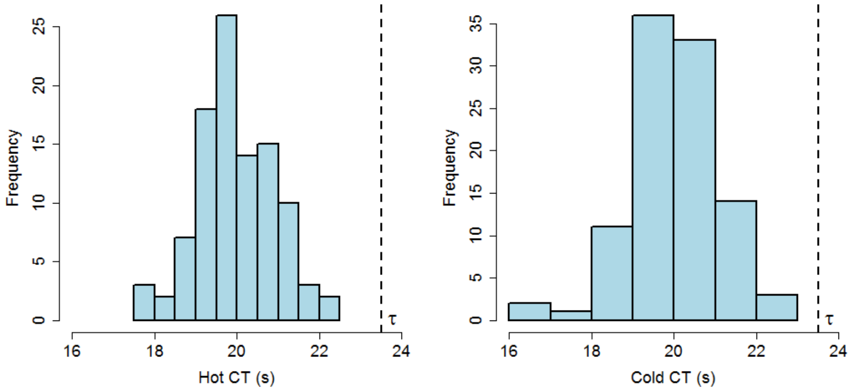

Figure 3.2: Histograms of the two simulated datasets, $n = 100$.

normal model, we have margin to the $\tau = 23.5s$ requirement (though not ample margin).

To determine whether the normal model is a good fit to these data, we use probability plots and goodness of fit tests, which are standard statistical tools for checking distributional fit (D'Agnostino and Stephens, 1986; Montgomery et al., 2009). Probability plots, often called quantile-quantile (QQ) plots, are simply plots of the estimated versus empirical quantiles of the data. The estimated quantiles are calculated under a statistical model (in this case, the normal distribution). If the model is correct, these plotted percentile pairs should form a line, within the uncertainty of the percentile estimates. Confidence intervals can be placed around this line to reflect uncertainty in the model fit. Probability plots for these data are displayed in Figure 3.3. For hot CT, the normal model appears to be a good fit for the data, For cold CT, there is some evidence of poor model fit in the lower tail of the data, though the tails are still within the point-wise confidence intervals; the lower-tail appears 'heavier' than would be predicted under the normal model (i.e. sample quantiles are greater than model predicted quantiles).

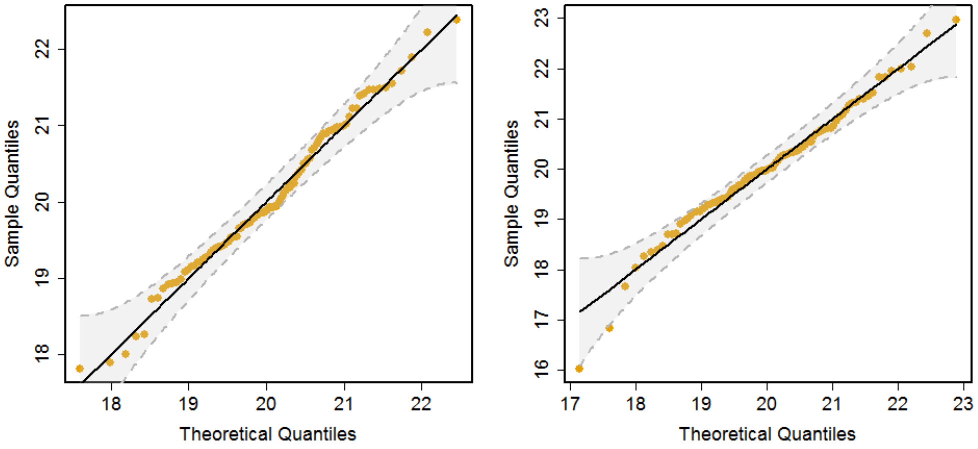

Figure 3.3: Probability plots for hot CT (left) and cold CT (right).

We then apply a distributional goodness of fit test to determine whether there is evidence of lack of fit for the normal model. Traditional statistics goodness of fit tests for distributional form (Stephens, 1974) are formulated based on the hypotheses:

H_0 : The model is a good fit for the data.
H_A : The model is not a good fit.

Differences between the empirical distribution of the data and the predicted distribution under the model are used to construct test statistics to provide evidence against the null (Stephens, 1974); the test statistics are then transformed into p-values and the null hypothesis is rejected when $p < .05$ or $p < .1$. A commonly-used goodness of fit test in QMU applications is the Anderson-Darling (AD) test (Razali et al., 2011), because this test weights the distribution tails more heavily than other goodness of fit tests (such as Shapiro-Wilk or Kolmogorov-Smirinov). Applying the AD test to our data, we test the null that the normal model is a good fit to the data versus the alternative that the normal is not a good fit. The p-value for hot CT is .19 and the p-value for cold CT is .22. Therefore, we fail to reject the null hypothesis and conclude that we do not have evidence of lack of model fit for either performance measure.

No evidence of lack of model fit is a critically different statement than evidence of model fit. Hence, this goodness of fit tests addresses the question, ‘Is there evidence that my parametric model is a bad fit to the observed data?’ However, this test does not address model validation for percentile estimation, which pertains to ‘Is the model sufficiently accurate for predicting tail behavior?’ Because these datasets were simulated, we know the true answer to this question. Hot CT was generated from a normal distribution, while cold CT was generated from a t distribution with 5 degrees of freedom (denoted t_5). These distributions are plotted in Figure 3.4, illustrating the heavier-tails of the t_5-distribution relative to the normal. The true 99.5^{th} percentiles for these distribution are $22.6s$ and $24.0s$. Hence, hot CT has margin to the $\tau = 23.5s$ requirement while cold CT does not. In fact, the failure rate for cold CT would be around 1%, doubling the .5% requirement. Underestimating the failure rate by a factor of 2 can have large implications when reliability estimates are rolled-up into system-level reliability models, and therefore the normal model was not an adequate approximation for predicting the 99.5^{th} percentile for cold CT.

3.1.3 New metrics

In Section 3.1.2, we highlighted the failure of goodness of fit tests and probability plots to effectively communicate uncertainty associated with the selected statistical model. We now propose alternative metrics to illustrate that, without knowledge of the underlying probability distribution, the credibility of percentile estimates depends: (1) how far out in the tails is the percentile of interest and (2) how much confidence about the estimate is required. When the intended use of the model is inferring tail behavior, we should consider:

1. Degree of extrapolation (Section 3.1.4): Is extrapolation outside the range of the observed data occurring?

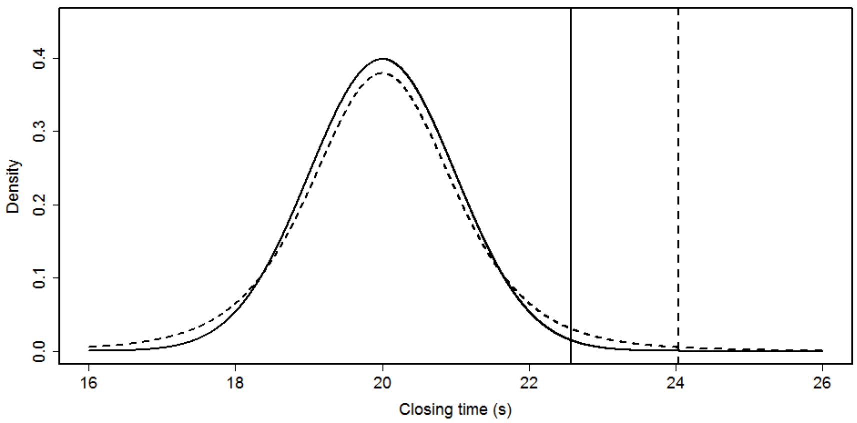

Figure 3.4: Two parametric distributions used as examples. Solid line is hot CT (normal distribution) and dashed line is cold CT (t_5 distribution); vertical lines correspond to true 99.5^{th} percentiles of these distributions.

2. Model fit in the tails (Section 3.1.5): How consistent are the observed tails of the data with the fitted model?

3. Sensitivity to model choice (Section 3.1.6): How much do the tail estimates change when the modeling assumptions are relaxed?

In the following subsections, we develop a suite of tools to answer these three questions.

3.1.4 Degree of extrapolation

First, we consider how far outside the range of the observed data we are extrapolating. To measure extrapolation, we simply compare the observed sample size n to the sample size that would be required for non-parametric tolerance interval construction (Equation 3.2):

$$\phi = \frac{log(\alpha)}{nlog(r)}. \tag{3.4}$$

where n is the sample size, r is the percentile of interest and $1 - \alpha$ is the desired confidence level. As seen in Figure 3.5, large sample sizes are required for full non-parametrics. This metric provides a useful representation of how much extrapolation is occurring when predicting to the tails. Further, the metric can also help the user understand the range of percentile estimates where extrapolation is not occurring.

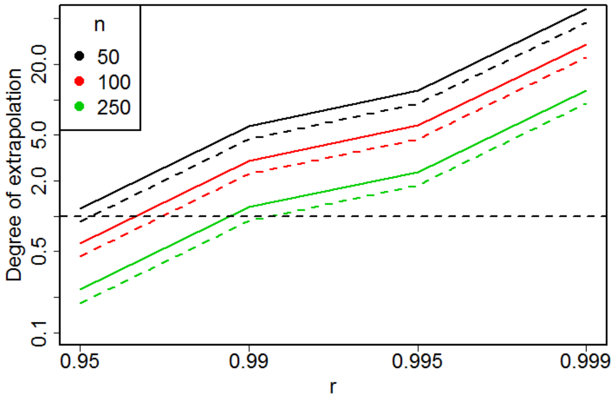

Figure 3.5: The non-parametric required sample size divided by the true sample size n as a function of the percentile of interest r at the 90% confidence level (dashed line) and 95% confidence level (solid line).

3.1.5 Model fit in the tails

To assess the model fit in the tails, we use return-level plots as a supplement to QQ plots. Return-level plots are a graphical tool from extreme value modeling (Coles et al., 2001) that examine the fit of a statistical model to the tails of the data. The return-level plot is simply a plot of the model-based percentile estimates and empirical percentiles as a function of the 'return-level,' denoted n_r for percentile r (Coles et al., 2001). The return-level n_r is defined as the number of units for which we would expect 1 failure to occur, $n_r = 1/(1-r)$. For instance, for a .99 reliability requirement, in a population of size 100, only 1 failure is allowed. This translates the 99% reliability requirement to a more intuitive '1 out of 100' failure rate, where the return level is $n_r = 100$. Typically, the return-level is plotted on the log-scale.

The return-level plot is constructed as follows:

- For order statistics $X_{(k)}$ where $k/n > .5$, plot the empirical quantiles as a function of n_r. Note that only the observations in the 'tail of interest' are plotted; i.e., for an upper requirement, only observations greater than the median are plotted.
- For $r > .5$, overlay a model fit-line using the predicted model-based quantiles as a function of n_r.
- Add the performance threshold τ and n_r corresponding to the requirement r as reference lines.

As with QQ plots, confidence intervals or posterior predictive intervals can be added to the return level plot to visualize uncertainty in the model fit. Using the asymptotic distribution of the order statistics, we can estimate the standard error which is used in turn to construct the confidence intervals(**?**).

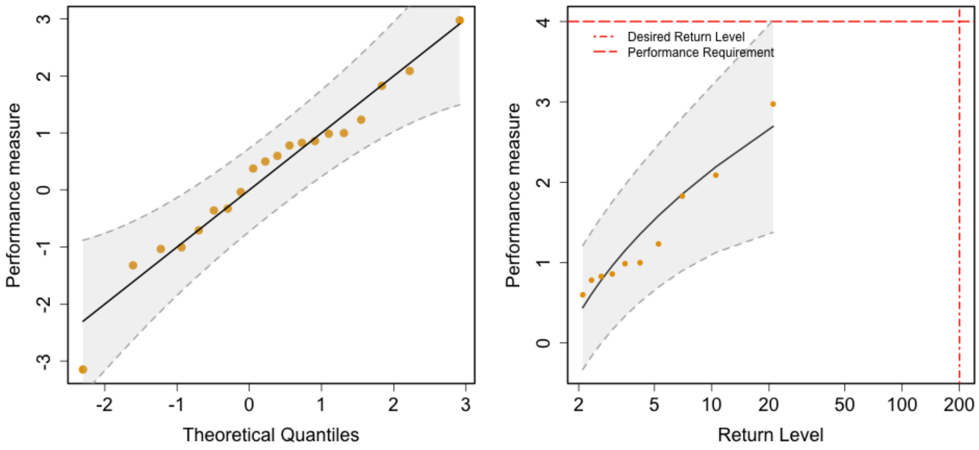

Figure 3.6: Simple example comparing a return-level plot (left) to a QQ plot (right) for tail extrapolation. A sample size of 20 is used to estimate the 99.5^{th} percentile to demonstrate margin to a requirement at 4. The extrapolation from a sample size of 20 to a population of size 200 is evidence from the return-level plot.

An example is shown in Figure 3.6. By mapping percentiles onto more intuitive population sizes, i.e. return levels, the lack of data in the tails with which to evaluate model fit is more apparent. On the other hand, QQ plots examines how well all of the observed data match a model-based prediction and are difficult to interpret for tail estimation (Scholz, 2005). Return-level plots are more appropriate for assessing how the model prediction performs near the target of estimation, an extreme percentile.

3.1.6 Sensitivity to modeling assumptions

Lastly, we construct a validation metric to provide evidence for our question of interest;"*Does the parametric distribution provide a sufficiently accurate estimate of the percentile of interest?*" To answer this question, we consider an approach in a similar vein as Diebolt et al. (2007), and compare fully-parametric models (Section 3.1.1) to more flexible extreme-value models (Section 3.1.1). Diebolt et al. (2007) constructs a goodness of fit test to compare a two-parameter parametric model to a GPD model, basing the test statistic on the difference between a percentile estimate under the two models. Following Diebolt et al. (2007), we also construct a metric based on the difference of a percentile under the two models.

Deviating from previous statistical approaches, we do not construct a goodness of fit test. That is, we are not aiming to show lack of parametric model fit, but rather provide evidence that the parametric model is a good fit, i.e. 'validate' the model. Specifically, we construct a validation metric aiming to show that a parametric model gives sufficiently similar percentile predictions as a more data-driven extreme-value model in the tails. This metric is similar in spirit to the 'reliability metric' used in the model validation literature, which provides a measure of the 'reliability,' or probability of success, of a model based on observed experimental data (Rebba and Mahadevan, 2008). In the computational simulation literature, many metrics for model validation have been developed to quantify model validity. (Mullins et al., 2016) differentiated two common categories of validation metrics, based on whether uncertainties are primarily epistemic or aleatory. In our application, the sources of uncertainty are both epistemic (sparse data and model form uncertainty), and therefore validation metrics targeted toward epistemic uncertainty are most appropriate (for a discussion of aleatoric metrics based on distributional shape comparison, such as the area metric, see (Ferson et al., 2008)). Epistemic uncertainty validation metrics, such as the reliability metric, aim to compare high probability regions between distributions (Rebba and Mahadevan, 2008). Because we are interested in determining similarity in two percentile estimates, where uncertainty is driven by epistemic uncertainty, a validation metric that compares regions of high probability between percentile estimates is most appropriate.

Definition of metric

We now define the proposed metric. Consider two models: $\mathscr{M}_1$, which uses only the data in the tails of the distribution (as in Section 3.1.1) and $\mathscr{M}_2$, which is a parametric model (as in Section 3.1.1). To assess the robustness of model $\mathscr{M}_2$, we examine whether $\mathscr{M}_2$ gives inferences that are sufficiently similar to the less parametric and more data-driven $\mathscr{M}_1$. We use Bayesian inference to construct a validation metric based on distances between the posterior distributions of the percentile estimates under the different models.

Given data X, let Q_r^k denote the estimated posterior distribution of percentile Q_r under model $\mathscr{M}_k$, for $k = 1,2$. For notational convenience, we suppress r and simply refer to Q^k and Q. Since $\mathscr{M}_1$ is more conservative (makes fewer parametric assumptions) than $\mathscr{M}_2$, then we would want to show that Q^2 is sufficiently close to or greater than Q^1 (for an upper requirement). A notional depiction of the percentile comparison approach is shown in Figure 3.7. If both models are correct, then, asymptotically, both models will provide estimates of Q that converge to the true value of Q.

Define $R_{\mathscr{M}_1,\mathscr{M}_2} = Q^1 - Q^2$, representing the difference in percentile estimates under $\mathscr{M}_1$ and $\mathscr{M}_2$. We can construct a posterior distribution on $R_{\mathscr{M}_1,\mathscr{M}_2}$ to make inferences about the reliability of $\mathscr{M}_2$ relative to the more flexible $\mathscr{M}_1$. To select model $\mathscr{M}_1$, we use the GPD model from Section 3.1.1, censoring the data at a point u and assume that $X - u \sim GPD$. The GPD model was selected to provide a flexible model for comparison to a fully parametric model using all collected data.

We use Bayesian inference (Gelman et al., 2014) to estimate the posterior distribution of Q^1, Q^2 and $R_{\mathscr{M}_1,\mathscr{M}_2}$ given a set of outcome observations $\boldsymbol{X}$. Bayesian inference is advantageous in this setting, due to the ability to construct a prior distribution that defaults to conservative percentile inferences with sparse data; frequentist inferential procedures, such as the bootstrap, could easily be substituted for estimation, but in sparse data situations, such frequentist procedures can underestimate uncertainty and should be used with caution (Schenker, 1985; Bergquist, 2006; Scholz, 2007).

Fitting the parametric model ($\mathscr{M}_2$) using Bayesian inference is straightforward. We specify non-informative priors on the model parameters and update the model parameters using Gibbs or Metropolis-within-Gibbs sampling, with the full process described in Appendix A. Fitting the GPD model is more involved. Specifically, to fit the GPD model ($\mathscr{M}_1$), we construct prior distributions such that, when data are limited, inferences about tail behavior will err on the conservative side. Specifically, we restrict the GPD model space to infinite tailed distributions by assuming $\xi \geq 0$; the parameter ξ governs the tail behavior of the GPD model. We place weakly informative empirical Bayesian priors on the model parameters to ensure model convergence; and also put a finite mass on $P(\xi = 0)$, which includes exponentially shaped right tails and encompasses many common distributions such as the normal, lognormal, Weibull, and gamma distributions. We sample from the posterior distribution of the model parameters using reversible jump MCMC. In appendix B, we describe details of the Bayesian estimation procedure for the GPD model, as well as results from a simulation describing the improved performance of the Bayesian GPD model relative to frequentist maximum likelihood estimation. We also provide R functions for implementing this metric as supplementary files. The posterior for $R_{\mathscr{M}_1,\mathscr{M}_2}$ can be obtained by sampling from the posteriors of Q^1, Q^2.

Inferences about the sensitivity of inferences to the selected model $\mathscr{M}_2$ can be based on summary measures from the posterior of $R_{\mathscr{M}_1,\mathscr{M}_2}$. We propose using $p_\varepsilon = P(R > \varepsilon)$ as the model sensitivity metric, where ε is a tolerance limit for the amount of acceptable error in the percentile estimate. Heuristically, p_ε can be interpreted as the confidence that $\mathscr{M}_2$ provides a percentile estimate that is within ε of the estimate under $\mathscr{M}_1$. Using this metric, we must show the model fit is adequate for percentile estimation, rather than look for evidence that the model is inadequate (as is the case with historical goodness of fit tests). The tolerance ε gives the user a 'buffer' for the amount of acceptable error in the percentile estimate.

Another potential metric is simply comparing the tolerance interval estimates under $\mathscr{M}_1$ and $\mathscr{M}_2$. If the difference between the tolerance interval estimates is sufficiently small, then we can conclude that the models provide sufficiently similar percentile uncertainty measures.

These validation metrics are designed such that, in order to conclude that there is evidence that the model is 'valid':

- Less data are required when we are willing to tolerate more error in the model predictions (specified through ε);
- More data are required as the target percentile moves further out in the tails of the distribution (p gets closer to 0 or 1), in order to compare model predictions to observed data in the space where prediction will occur.

Neither of these desirable properties hold for distributional goodness of fit tests.

Performance of model sensitivity metric

We conducted a simulation study to assess how p_ε changes as a function of the percentile r and ε, assuming the underlying model for the data is known. We consider the same two data generating

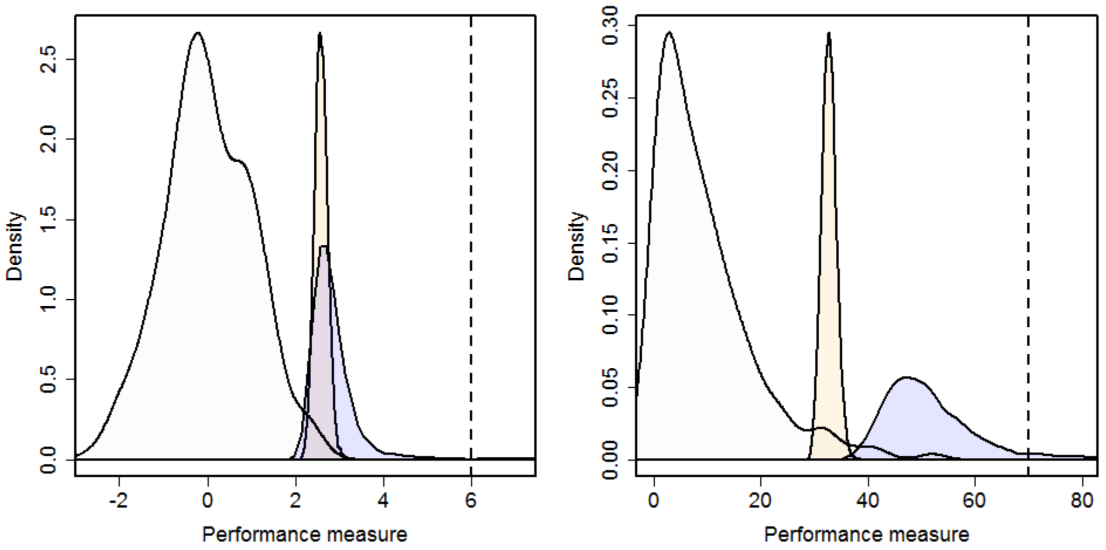

Figure 3.7: Posterior distributions for Q^1 under parametric model (orange) and Q^2 under more robust model (purple) for a 99.5^{th} percentile estimate based on $n = 100$ observations when the model is correct (left) and incorrect (right). A kernel density estimate of the data distribution is show in grey. (Left) $\mathscr{M}_1$ is correct and percentile estimates are very similar, but the robust model has more uncertainty. (Right) $\mathscr{M}_1$ is incorrect, the less parametric model $\mathscr{M}_1$ gives a much higher estimate of the percentile.

mechanisms that were used for the launch safety device example in Section 3.1.2: a normal distribution and t_5 distribution. We compare the fit of the normal model $\mathscr{M}_2$ to the GPD model with 10% of the data censored. We simulated 1,000 different datasets of size $n = 100$ and estimated p_ε for $\varepsilon = \{0, .1, ..., 2\}$ and for $p = \{.95, .99, .995, .999\}$.

In Figure 3.8, we plot the average p_ε over the 500 simulations as a function of ε, as well as the probability that $p_\varepsilon > .9$ as a function of ε. The metric p_ε increases as ε increases and as p decreases. When the data are generated from the t_5 model, ε must be large ($>> 2$) to bound the GPD model with the normal model $+\varepsilon$ (with 90% confidence) when $r > .95$ (note that the 95^{th} percentile can be estimated from the $n = 100$ samples without extrapolation and that the t_5 and normal 95^{th} percentiles are quite similar). The conservatism in the GPD model is evident and is expected, given that priors for the GPD model were defined to default to more heavy-tailed distributions when data are limited.

3.2 Results

In this section, we apply the proposed tools to the launch safety device QMU example from Section 3.1.2. Recall that we are aiming to estimate the 99.5^{th} percentile for closing time, and calculate corresponding 95% tolerance intervals, with the goal of demonstrating margin to a $\tau = 23.5s$ requirement. We now apply the validation metrics detailed in Sections 3.1.3; specifically, we calcu-

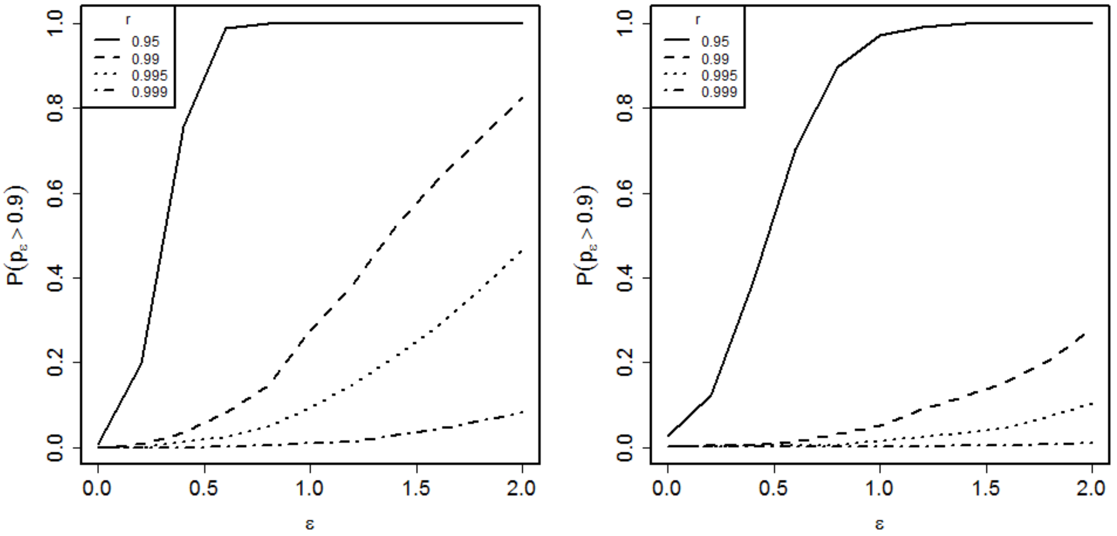

Figure 3.8: The probability of p_ε exceeding .9 as a function of ε when: (left) the normal model is correct and (right) the t_5 model is correct.

late the degree of extrapolation, examine model fit in the tails, and consider changes in percentile estimation under the more robust tail-based model.

Degree of extrapolation. When $n = 100$ and $\alpha = .05$ (95% confidence), extrapolation is occurring beyond the 97^{th} percentile (Figure 3.5). To estimate the 99.5^{th} percentile with 95% confidence, as in the Section 3.1.2 example, we would need $n = 598$ and hence a $\phi = 6$ times larger sample, suggesting a somewhat high degree of extrapolation.

Model fit in the tails. Return level plots for closing time are shown in Figure 3.9. For both hot and cold CT, the normal model does not appear unreasonable for the data. However, unlike the QQ plot, the return-level plot highlights the fact that we are using 100 units to predict performance of $n_r = 200$ units and hence inferences about the 99.5^{th} percentile reflect extrapolation outside the range of the data.

Validation metrics. We estimated the posterior distributions for the 99.5^{th} percentiles of hot and cold CT under the normal model ($\mathscr{M}_2$) and the GPD ($\mathscr{M}_1$). For the GPD model, we chose u such that only the upper 10% of the data were used to fit the model. We then estimated the posterior distribution for $R_{\mathscr{M}_1,\mathscr{M}_2}$ and calculated p_ε for various values of ε.

First, we summarize the results for hot CT. The fitted GPD and normal models give similar 99.5^{th} percentile point estimates, with posterior median 22.5s for both models (Figure 3.10). However, uncertainty is higher under the GPD model, as demonstrated through the model validation metrics (Figure 3.11). The metric p_ε exceeds .8 when $\varepsilon > 1$ and exceeds .9 when $\varepsilon > 2.5$. The difference in tolerance interval estimates increases as the confidence level increases, as expected, highlighting the increased uncertainty as the degree of extrapolation increases. The 99.5/95% tolerance interval for the GPD model is 26.1s, approximately 3.2s higher than under the normal model.

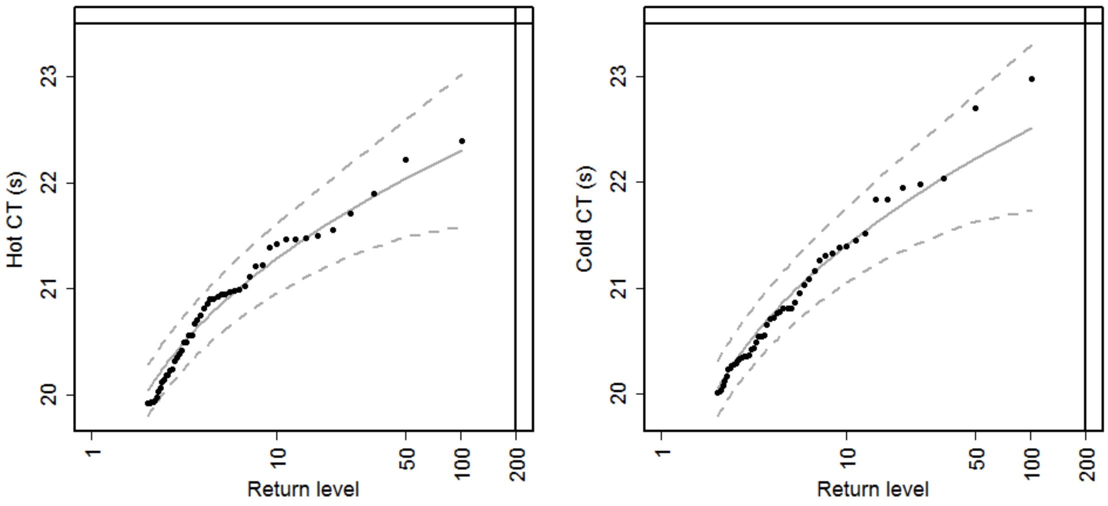

Figure 3.9: Return-level plot for hot CT (left) and cold CT (right). In the return-level plot, the x-axis is on the log-scale. The vertical line represents the reliability population size n_r corresponding to $r = .995$. The horizontal line is the performance threshold $\tau = 23.5s$. The black dots are the observed quantiles as a function of the reliability population size; the blue lines are the theoretical quantiles (solid) with 90% confidence bounds (dashed).

For cold CT, the GPD model provides more conservative inferences, with posterior median of the 99.5^{th} percentile at 23.4s compared to 22.9s under the normal model. As in the hot CT case, uncertainty is higher under the GPD model (Figure 3.11). The metric p_ε is lower than the hot CT case for lower values of ε. Further, the difference between tolerance interval estimates is larger in the cold CT case, until the confidence level is sufficiently high, as anticipated. The 99.5/95% tolerance interval for the GPD model is $26.1s$, approximately 2.7s higher than under the normal model.

For both hot and cold CT, the GPD-based percentile estimate has more uncertainty, due to the fact that this model uses only the tails to estimate the percentile and makes few assumptions about the shape of the tail (the normal model assumes exponentially decaying tails). The model validation metrics highlight the impact of the GPD uncertainty (Figure 3.11). The standard deviation for both hot and cold CT is close to 1, and percentile inferences could change by 1-3 distribution standard deviations under a different statistical model (depending on the user-selected threshold for p_ε or confidence level for the tolerance interval). Hence, considering that there is limited margin in this example, the percentile estimates are rather sensitive to the model form assumptions.

Summary. Combining these three validation metrics, we conclude that: we are extrapolating outside the range of the data, the model fit cannot be evaluated where prediction will occur, and the percentile estimates are somewhat sensitive to selection of the normal model. Hence, we can extrapolate to estimate the 99.5^{th} percentile and a corresponding tolerance interval, but we should not be surprised by model form error invalidating our statistical inferences. Note the difference in information provided from these metrics as compared to the more traditional probability plots and

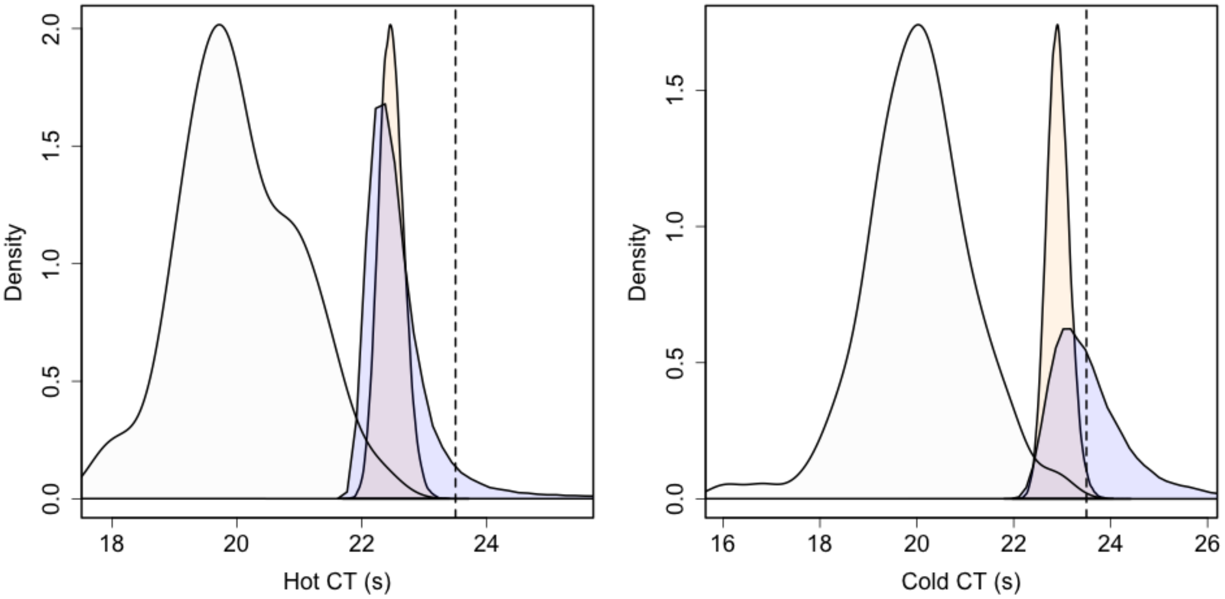

Figure 3.10: Kernel density estimates of the posterior distributions for the 99.5^{th} percentile. Data were generated from the normal distribution (left) and t_5 distribution (right). Percentile estimates were calculated under the normal model (orange) and the GPD model censoring the lower 90% of the observations (blue). The vertical dashed line reflects the hypothetical threshold of $\tau = 23.5s$. The grey kernel density estimate of the data is also shown.

goodness of fit tests (Section 3.1.2).

3.3 Discussion

The proposed tools provide a direct method for communicating the irreducible risk associated with tail extrapolation in low probability, high consequence engineering applications. Clear communication and/or visualization of risk is essential for interpreting and utilizing statistical analyses (Spiegelhalter et al., 2011). We demonstrated the inadequacy of distributional goodness of fit tests and probability plots for model validation in tail estimation and proposed a set of new statistical tools to achieve this objective. We emphasize that statistical model validation and computation simulation model validation are not conceptually or technically different tasks. Future work should aim to highlight the notion that these two validation activities do not need to be viewed differently. By relating the proposed metrics back to the notion of validation in computational simulation, these metrics may be more intuitive to engineers and analysts who are familiar with validation in this context.

Our illustrative example pertained to QMU for nuclear weapon stockpile evaluation. Multiple different representations of uncertainty for QMU have been proposed in the literature, including probability theory (enveloping both Bayesian and frequentist statistical inference), evidence theory,

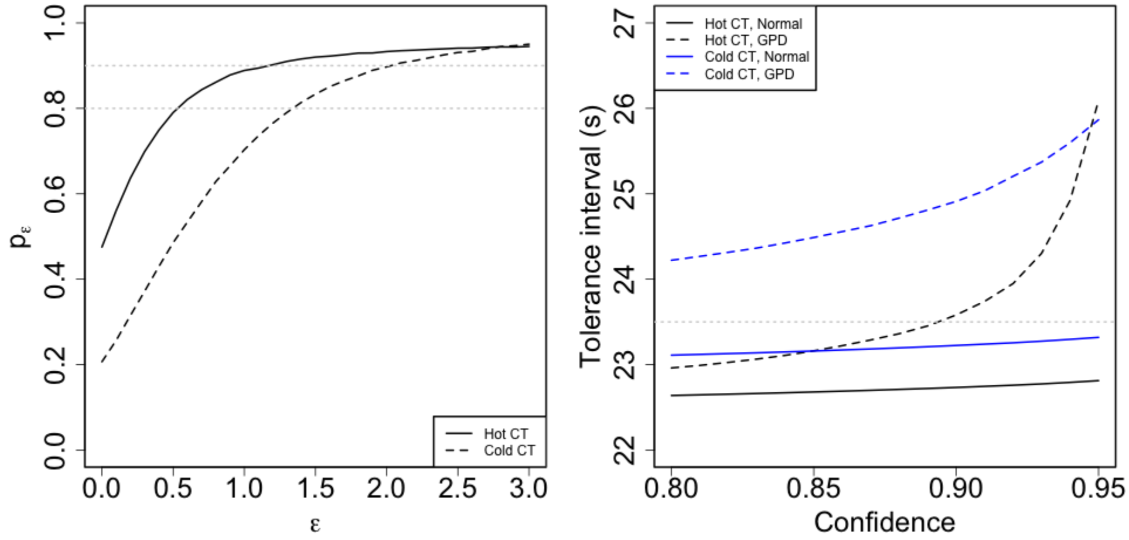

Figure 3.11: Validation metrics. (Left) p_ε for hot and cold CT performance measures. (Right) Tolerance interval estimates for 99.5^{th} percentile as a function of the confidence level for hot and cold CT performance measures under the two different models.

and possibility theory (Helton, 2011). While substantial efforts have been made to delineate different potential inferential frameworks for QMU, little effort has been placed on providing guidance for choosing between these frameworks. In practice, when experimental data are available, probabilistic modeling is used in the vast majority of QMU applications without explicitly considering risk associated with model misspecification.

Parametric distribution fitting using two-parameter distributions (Section 3.1.1) is prevalent in many engineering applications. We postulate that reasons for popularity include ease of estimating the distribution, ease of using the results, and lack of data availability for less parametric approaches. More flexible parametric approaches could be considered, such as four-parameter distributions (Yeo and Johnson, 2000; Su, 2009), mixture distributions (Roeder and Wasserman, 1997), and non-parametric estimates (Pradlwarter and Schuëller, 2008). The additional flexibility of these approaches will result in better accuracy in percentile estimation over two-parameter distributions, at the cost of efficiency. The model validation metrics proposed in this paper can also be applied to these more flexible distributions.

The purpose of this paper was to provide a novel framework for evaluating parametric distribution fits when these distributions are used for tail estimation. The proposed methods have several limitations. First, the scope of the examples in Section 3.2 was limited to two restrictive examples for sake of brevity. In our example, the normal model approximation resulted in anti-conservative percentile estimates. There are many other examples that could be considered where the normal model is a conservative approximation (Romero et al., 2013). Future work could consider testing the validation metrics on a broader set of candidate problems. Additionally, to assess model sensitivity, we developed a novel metric to compare parametric models to extreme-value models using only observations in the tail. The GPD model produced very conservative inferences, due to the conservative prior parameterization that we selected for the GPD model. While this conservatism is desirable in high-consequence applications, future work could explore alternatives to the GPD model. As an example, we could have compared the normal model to a censored version of the normal model, where, similar to the GPD approach, observations are censored below a threshold u. Fitting tail-based models requires a sufficient amount of data to be able to censor observations at u and fit a model to the uncensored data. If the data are too sparse to be able to fit such models, then extreme percentile estimation is likely not a prudent objective without an underlying physical model for the outcome.

Estimating rare events using limited data is always a challenging task; our goal herein was the development of concise metrics to articulate the risk of model form misspecification and extrapolation. We hope that use of these validation metrics in practice will help improve risk communication and ultimately improve the process for characterizing rare events using statistical inference.

Chapter 4

Propose information integration models

Traditional tolerance intervals based on frequentist statistics allow for confidence and coverage levels to be specified and compensate for the available quantity of data. We demonstrate in section 2.2 why the application of traditional tolerance intervals in sparse data situations can easily lead to misinforming results. One means of reducing the risk when conducting QMU analyses in sparse data situations is to better leverage all available knowledge about the problem of interest. Tolerance intervals based on Bayesian statistics offer a means of calculating tolerance intervals, while also incorporating expert knowledge into the estimates. Bayesian tolerance interval estimation is not a new statistical technique (Aitchison, 1964; Krishnamoorthy and Mathew, 2009), but their utility in integrating information to make conservative and useful estimates of extreme percentiles could not be located in the existing literature. In this section, we consider how Bayesian hierarchical modeling might improve percentile estimation in QMU applications.

4.1 Methods

Bayesian tolerance bounds can be calculated using Bayesian hierarchical modeling. In the present context, Bayesian hierarchical modeling refers to modeling the uncertainty a distribution's hyper-parameters instead of the uncertainty in the distribution. Once prior knowledge is used to specify a distributional form for the data as well as the hyper-parameter prior distributions, the uncertainty in the hyper-parameters can then be calibrated to the available data. Propagating the hyper-parameter posterior distribution through the specified distributional form will result in a family of distributions (of the specified distributional form), from which tolerance bound estimates can be made. Expert knowledge about attributes of the underlying distributional form can be incorporated into the assumed distributional form model that then impacts the likelihood function or into the prior distributions of the hyper-parameters of the assumed distributional form. For instance, if a tail subpopulation like that previously shown in Figure 1.5 was known to exist, the assumed distributional form could be formulated as a mixture of normal distributions. For that mixture of normal distributions, the means, standard deviations, and relative weighting of the distributions would be the hyper-parameter's whose uncertainty would be calibrated to the available data. Additional information about the likely location and scale of the subpopulation could be incorporated into prior distributions of the second distribution's mean, standard deviation, and relative weighting.

Bayesian networks are a helpful method for visualizing Bayesian hierarchical systems (Mahadevan and Rebba, 2005; Urbina et al., 2012). Bayesian network representations of a T distribution and a normal distribution with a tail subpopulation are shown in Figure 4.1. For both networks the observed data d_i is directly comparable to the assumed distributional forms D. Based on our assumed model form (M) for D, information about the hyper-parameters θ can be learned

$$P(\theta|D=d_i,M) \approx P(D=d_i|\theta,M)P(\theta). \tag{4.1}$$

Visualizing problems with Bayesian networks is helpful for checking the solution's mathematical construction as well as communicating the approach.

As an example application of Bayesian tolerance intervals estimation, consider the normal mixture model Figure 4.1. If there is a subpopulation in the data, but we have not sampled from the subpopulation, prior distributions could be placed on the subpopulation frequency w, mean μ_2, and variance σ_2 to inform how inferences change in the presence of the subpopulation. Of course, results will be highly sensitive to the priors, and therefore the priors should be informed by a relevant source of data, such as a computational simulation model or a similar component (e.g. similar design on a different system or previous build for development components).

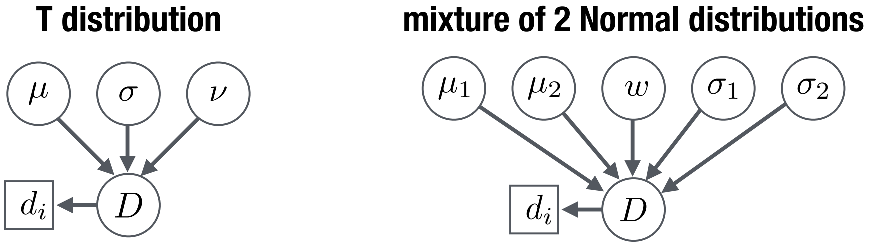

Figure 4.1: Bayesian network representations of Bayesian hierarchical models used to determine tolerance intervals. The left network represent a T distribution with parameters μ for location, σ for scale, and ν for the degrees of freedom. The right network represents a mixture of two normal distributions, where w is the relative weighting of the two distributions.

4.2 Results

For the first illustration of Bayesian tolerance intervals, we assume that we know that the data comes from a T distribution with significant tails approximately characterized by a T_5 distribution. The prior distributions used for the location and scaling of the distribution are $\mathscr{U}(-2,2)$ and $\mathscr{U}(0,3)$. The uncertainty in the location and shape scaling of that distribution are then explored using a MCMC implementation within the software package STAN, accessed through the Python interface PyStan (Stan Development Team, 2016 Version 2.14.0.0). Only 4,000 random combinations of samples are used to determine the empirical reliability for each tolerance bound due to the computational expense associated with each Bayesian computation. Figure 4.2 and 4.3 show how Bayesian tolerance intervals for this T_5 distribution scale with the quantity of available data, for different confidence and coverage standards.

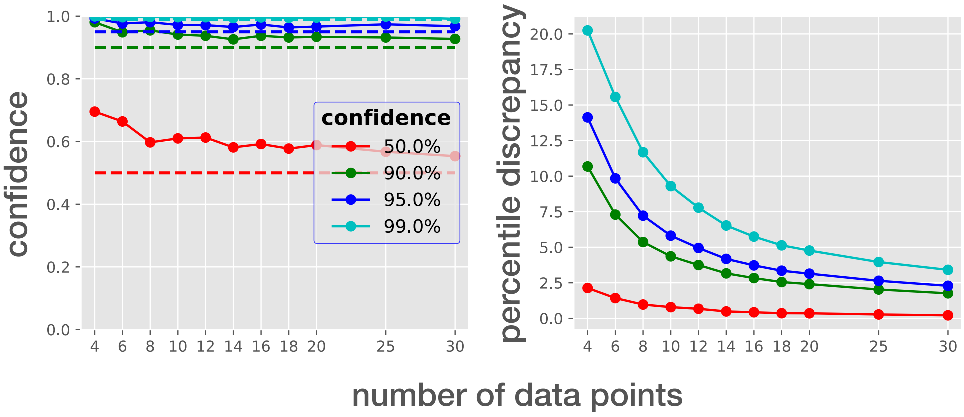

Figure 4.2: How the empirical reliability (left plot) and percentile discrepancy (right plot) of the 99.9% coverage Bayesian tolerance intervals applied to data from a T_5 distribution scales with the number of data samples. Trends depending on different prescribed confidence levels (50%, 90%, 95%, and 99%) are shown as solid lines with dots and corresponding prescribed confidence levels are shown as dashed lines. Values shown are average values from 4,000 random combinations. Bayesian tolerance intervals explore the uncertainty of the distribution's location and scale of the T distribution, but assumed the degrees of freedom are known.

The empirical reliability of all confidence levels considered for the 99.9% coverage tolerance bound are conservative, and this measure appears to scale towards the desired confidence level as the number of data points increases. In this case, larger quantities of data cause the posterior distribution to converge from the flat prior uncertainties to uncertainty distributions with high densities around the true values. The percentile discrepancy observed for small data quantities is sensitive to the prior distributions. Greater uncertainty in the hyper-parameter priors would result in larger percentile discrepancy for small sample sizes, but will have less impact for larger sample sizes whose posterior distributions are more influenced by the likelihood.

The 4,000 combinations appears to have been insufficient to smooth and separate the empirical reliability of the different coverage levels for the 95% confidence tolerance bounds. All coverage levels considered remain conservative and appear to scale towards the specified confidence level as larger sample sizes are used. The percentile discrepancy remains positive and scale towards zero, indicating that with enough data the percentile estimates should asymptote to the true value.

Next, we try the more challenging and realistic problem of estimating data from a T_5 distribution with uncertainty about its location, scale and degrees of freedom. When dealing with sparse data and increasing numbers of uncertain parameters, more informative hyper-parameter priors will be helpful. For this study the location prior is $\mathcal{N}(0,3)$, scale prior is $\mathcal{U}(0,3)$ and degrees of freedom prior is $\Gamma(2,0.3)+1$. Prior specification for the degrees of freedom hyper-parameter requires additional study. In the context of tolerance interval estimation, it is desired that the priors force conservative tolerance interval estimates in sparse data situations; see Section 3.1.6 for an example. Placing uniform uncertainty from 1 to ∞ distributes too much prior density to light-tailed distributions. A prior distribution that places more weight on heavier tailed distributions allows

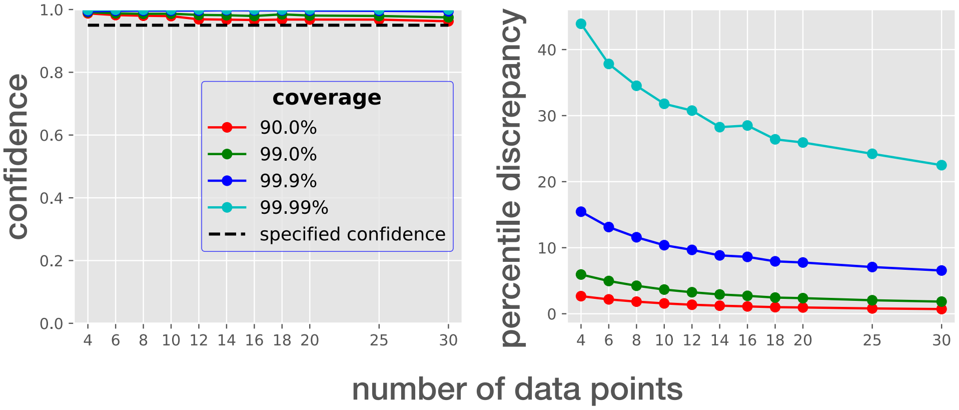

Figure 4.3: How the observed reliability (left plot) and percentile discrepancy (right plot) of the 95% confidence Bayesian tolerance intervals applied to data from a T_5 distribution scales with the number of data samples. Trends depending on different prescribed coverage levels (90%, 99%, 99.9%, and 99.99%) are shown as solid lines with dots and corresponding prescribed confidence levels are shown as dashed lines. Values shown are average values from 4,000 random combinations. Bayesian tolerance intervals explore the uncertainty of the distribution's location and scale of the T distribution, but assumed the degrees of freedom are known.

for tolerance interval estimates for sparse data to provide estimates that remain conservative by assuming thick tailed distributions until sufficient data proves otherwise. The gamma distribution prior on the degrees of freedom parameter attempts to favor heavy tails and was inspired by Juárez and Steel (2010). Figure 4.4 and 4.5 show how Bayesian tolerance intervals based on uncertainty in the location, scale, and degrees of freedom of a T distribution scale with different amounts of coverage, confidence, and data in a dataset.

The first item of note in the performance of the Bayesian tolerance intervals is that the 99.9/50 TIs quickly become less conservative as larger datasets are used. This is likely due to the prior specification for the degrees of freedom not placing enough weight on thicker tailed T distributions, but the percentile discrepancies show that the estimates remain close to the true values. Although this performance for low confidence levels is undesirable, as noted in the scaling study, 50% confidence is never a prudent choice for a the confidence level. On the other hand, the three higher confidence level-based TIs are conservative and appear to be scaling towards to the true values. With larger datasets, we hypothesize that the lower confidence level TIs will trend back towards to true percentile values as the posterior distribution converges to the true values.

The performance of the Bayesian tolerance intervals for different coverages are all conservative, likely due to being based on 95% confidence. All coverage levels appear to trend towards the specified confidence level and true percentile value.

With Bayesian tolerance intervals, prior specification has a significant impact for sparse data situations. If significant prior information is available, accurate and conservative estimates should be viable even with sparse data. If minimal prior information is available, conservative estimates can be made if the Bayesian problem is formulated with that goal. Inferences will be biased when

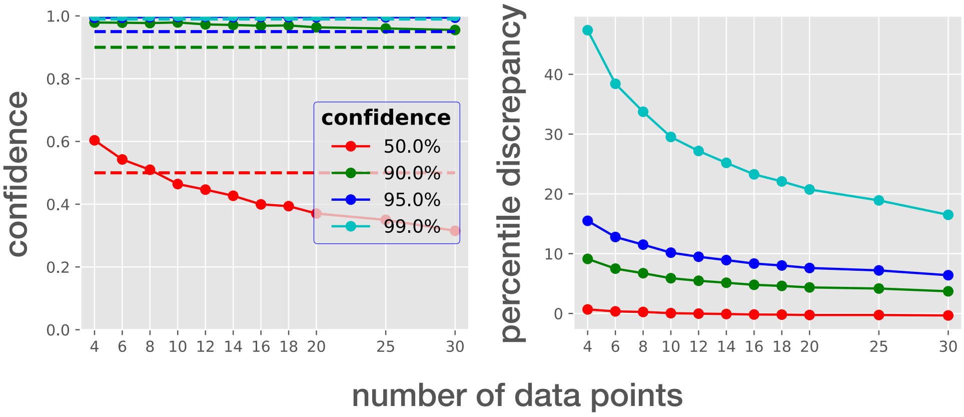

Figure 4.4: How the observed reliability (left plot) and percentile discrepancy (right plot) of the 99.9 coverage Bayesian tolerance intervals applied to data from a T_5 distribution scales with the number of data samples. Trends depending on different prescribed confidence levels (50%, 90%, 95%, and 99%) are shown as solid lines with dots and corresponding prescribed confidence levels are shown as dashed lines. Values shown are average values from 4,000 random combinations. Bayesian tolerance intervals explore the uncertainty of the distribution's location, scale, and degrees of freedom.

incorrect prior information is utilized, but incorrect user-information (such as model choice) will also bias non-Bayesian tolerance interval estimates, as see in Section 2. When applied to the aforementioned example problems, conservative estimates were found for all confidence levels that are suggested as prudent choices for QMU analyses.

4.3 Discussion

The integration of information from multiple sources appears to be one promising method of dealing with data poor environments. However, the exact means of implementing this framework in QMU applications are not well-established and involve the injection of user specific information, which can lead to bias if this information is incorrect. For instance, when using the Bayesian hierarchical modeling approach, it is left to the modeler to chose which data will be used and how that data is leveraged. The greater agility of the Bayesian hierarchical modeling approach makes it difficult to provide concise processes for its application, but the flexibility also allows the approach to better adapt to the large variability found in applications. The best advise for applying Bayesian hierarchical modeling to QMU problems is to transparently communicate modeling decisions, so that the bias is understood. For instance, it was found in creating the results shown in section 4.2 that the tolerance interval estimates were sensitive to the prior specification and that the sensitivity was correlated with the amount of data available. This is not surprising for Bayesian methods, where it is known that prior information dominates the posterior until sufficient evidence is available to surpass the prior understanding's impact. What was gained from applying the Bayesian

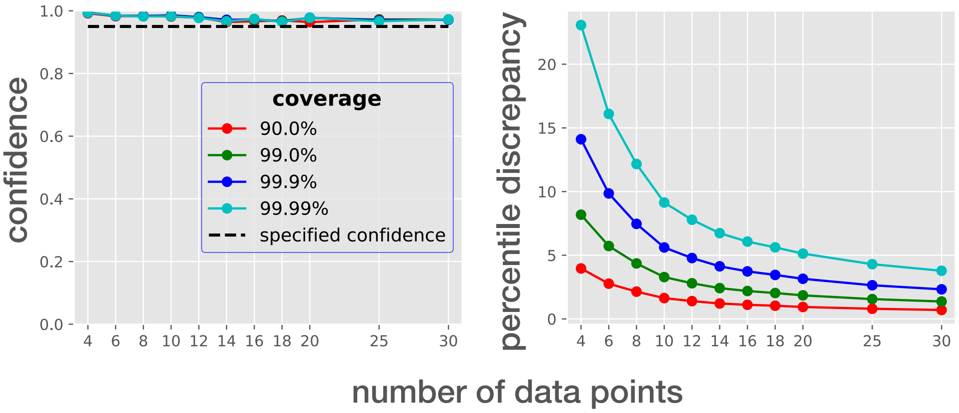

Figure 4.5: How the observed reliability (left plot) and percentile discrepancy (right plot) of the 95% confidence Bayesian tolerance intervals applied to data from a T_5 distribution scales with the number of data samples. Trends depending on different prescribed coverage levels (90%, 99%, 99.9%, and 99.99%) are shown as solid lines with dots and corresponding prescribed confidence levels are shown as dashed lines. Values shown are average values from 4,000 random combinations. Bayesian tolerance intervals explore the uncertainty of the distribution's location, scale, and degrees of freedom.

hierarchical modeling approach to T_5 distribution based data was a relatively simple approach to estimating tolerance intervals for "normal appearing" data with potentially greater weight in the tails. Classically, this data would have been assumed to be normally distributed, and hypothesis tests would be unlikely to provide evidence against that hypotheses, leading to non-conservative tolerance interval estimates, as was demonstrated in subsection 2.2.3. The cost of using this framework was the need to communicate the sensitivity of the estimates to any prior information used and the greater amount of knowledge it takes to apply the framework, as compared to the application of classical tolerance intervals. Additional exploration and application of Bayesian hierarchical modeling for QMU applications could reduce the cost of the approach, through increasing the community's knowledge base on the approach and streamlining methods of communicating user bias when results are presented.

Chapter 5

Anctipated Impact

The results in this report were generated from a one-year exploratory express (ExEx) LDRD project funded by the Engineering Sciences Investment Area. In this ExEx LDRD, we quantified the large, irreducible risk associated with current practices for experimental QMU. Specifically, statistical model-form uncertainty is currently ignored when calculating confidence in margin assessments, but this source of epistemic uncertainty typically dominates other sources of uncertainty. This project effectively highlighted this unquantified risk; however, there remains a need to update QMU methodologies to mitigate this risk, adapting to challenges of monitoring the reliability and safety of the NW stockpile in the 21st century. This project effectively laid the groundwork for researching new methods to improve QMU for experimental data at SNL. We illustrated the problem, developed metrics for communicating risk, and proposed a path forward for using data integration to mitigate risk.

This ExEx project should have high impact for a relatively small-scale project. We have already secured follow-on programmatic funding, are actively working to transition the methods into the mission-space, and have submitted a follow-on full LDRD proposal to the Engineering Sciences investment area. We feel that QMU should be an active research area for the labs and are optimistic that this project can generate some interest in this research area.

Follow-on funding. Results from this project are already having an impact on the SNL mission. Our team has procured programmatic funding to revise the QMU handbook and QMU trainings during FY17. This funding was a direct result of this LDRD work. Within the handbook revision, we aim to downplay the role of the tolerance intervals and formal statistical inference in sparse data situations (e.g. development), where heavy extrapolation is required. We will remove distributional goodness of fit tests from the handbook, based on the results from this project, and highlight challenges associated with selecting a model form for tolerance interval estimation. Further, we plan to remove methods that encourage heavy extrapolation outside the range of the data, such as predicting aging trends into the future based on an unvalidated linear model. We will also include information surrounding how computational simulation modeling can be leveraged in experimental QMU applications; specifically, we aim to emphasize that experimental and computational simulation QMU cannot be treated as distinct entities (which is common in practice) if we aim to quantify margin and uncertainty as accurately as possible. The handbook revision will highlight the fact that, in QMU applications, uncertainties are high-dimensional and difficult to quantify. Subsequently, in practice, data must be leveraged from multiple sources to fully characterize risk. In short, QMU is an engineering exercise that sometimes uses statistics, not a statistical analysis that sometimes uses engineering.

Mission impact. The Statistical Sciences group works closely with component PRTs in development to conduct QMU analyses to demonstrate positive margin in development, as well as with surveillance analysts to conduct and review QMU analyses for Annual Assessment Reviews. We are currently partnering with the NW development programs to use the results of this project to improve upon margin assessment strategies in development.

New proposals. We received funding for a full, 2-year LDRD for FY18-20 to build on the work developed within this project. The goal of this submitted proposal is to develop a unified Bayesian modeling framework for experimental (i.e. physical simulation) and CompSim QMU. We hypothesize that the proposed innovations will improve the robustness and credibility of margin assessments. The proposed project represents a paradigm shift for experimental QMU, moving toward a framework where margin is estimated by integrating all available information in a risk-informed fashion, rather than extrapolating from un-validated and non-robust models; we demonstrated the need for this paradigm shift in this project.

Chapter 6

Conclusions

The goal of this project was to illustrate challenges in statistical estimation of percentiles for quantification of margins and uncertainty (QMU) with sparse experimental data. To achieve this goal, we addressed three objectives:

1. **Quantify model form risk:** Quantify the consequences of model misspecification using a scaling study.
2. **Define model validation metrics:** Develop improved model validation metrics.
3. **Propose information integration models:** Explore potential solutions for relaxing stringent model form assumptions through integrating multiple data sources.

In the first objective, we demonstrated that tolerance intervals are not robust to the underlying distributional form. In the second objective, we proposed a new approach to validating probability distribution models that highlights degree of extrapolation, the model fit in the tails, and the sensitivity of estimates to the selected model form. In the third objective, we outlined a proposed method for integrating more information into tolerance interval estimates using Bayesian modeling.

By completing these three objectives, we demonstrated a need for new recommendations for QMU with experimental data. Data are never exactly characterized by parametric probability distributions, and distributional form uncertainty can compromise the credibility of tolerance interval estimates. We also illustrated how this problem is exacerbated by the fact that statistical tools for probability distribution evaluation provide uninterpretable and misleading information, but are used systematically due to a lack of accessible alternatives. This preliminary work will contribute toward our ultimate goal of developing a workflow for distributionally robust QMU.

In this project, we have illustrated that decision-making using QMU is a complex process that cannot be achieved using statistical analyses alone. Current statistical approaches can introduce risk through extrapolative percentile estimation. While there is a clear advantage to having simple, algorithmic statistical methods for QMU to encourage implementation, the cost of simplicity is often credibility. In Figure 6.1, we present a flow chart for assessing the credibility of statistical methods. In practice, an emphasis must be made on determining and acknowledging when there is not enough information to conduct a principled statistical analysis; introducing unverifiable modeling assumptions to increase information compromises credibility. Experimental data and statistical analysis have substantial value in informing the safety and reliability of our NW stockpile; however, understanding the limits of the statistical methods is necessary to avoid overconfidence in results and compromise credibility in data-driven inferences. We hypothesize that data integration, i.e. combining sources of information, to improve credibility is the future of data-driven QMU.

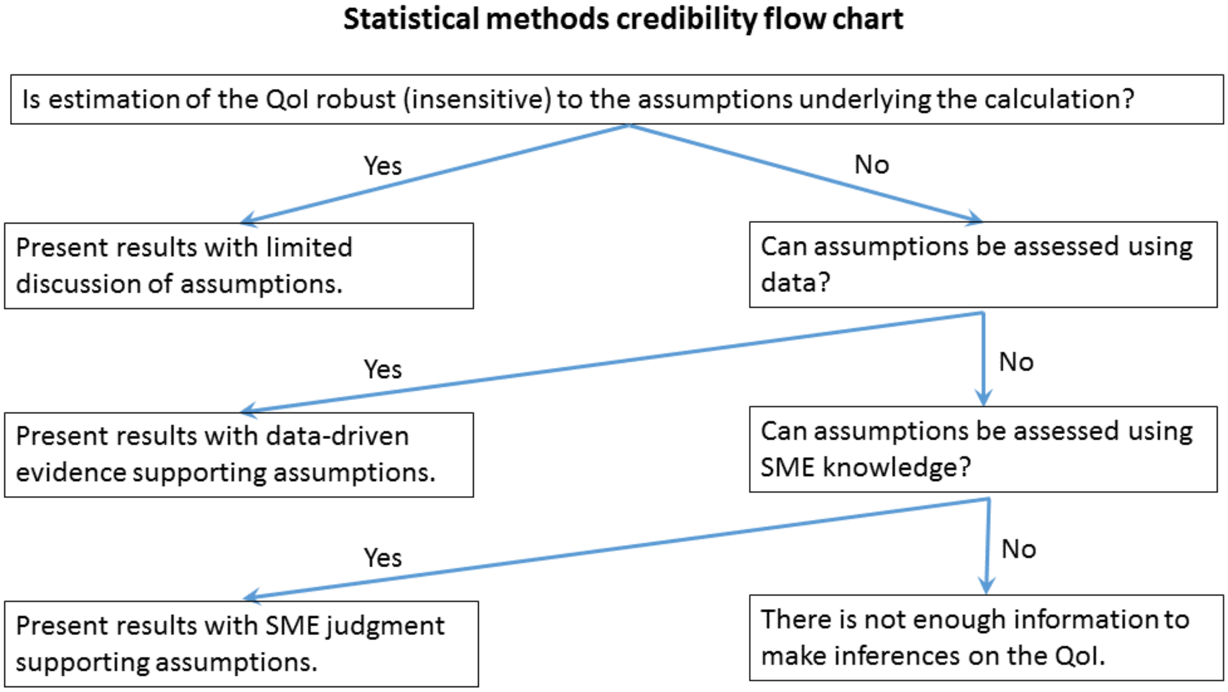
Statistical methods credibility flow chart
Is estimation of the QoI robust (insensitive) to the assumptions underlying the calculation?
Yes
No
Present results with limited discussion of assumptions.
Can assumptions be assessed using data?
Yes
No
Present results with data-driven evidence supporting assumptions.
Can assumptions be assessed using SME knowledge?
Yes
No
Present results with SME judgment supporting assumptions.
There is not enough information to make inferences on the QoI.

Figure 6.1: Flow chart to decide whether and how to present statistical methods for decision-making.

Appendix A: Bayesian estimation for the normal and normal-censored models

We use the normal-inverse gamma model for fitting the normal distribution to the data X:

$$\begin{aligned} X &\sim N(\mu, 1/\tau) \\ \mu &\sim N(\mu_0, n_0\tau) \\ \tau &\sim Ga(\alpha, \beta) \end{aligned} \tag{6.1}$$

which results in posterior:

$$\begin{aligned} \mu|\tau, X &\sim N\left(\frac{n\tau\bar{x} + n_0\tau\mu_0}{n\tau + n_0\tau}, (n\tau + n_0\tau)^{-1}\right) \\ \tau|X &\sim Ga\left(\alpha + \frac{n}{2}, \beta + \frac{\sum(x-\bar{x})^2}{2} + \frac{nn_0}{2(n+n_0)(\bar{x}-\mu_0)^2}\right) \end{aligned} \tag{6.2}$$

For the priors, we select $n_0 = 10^{-4}$ and $\alpha = \beta = .01$. Inferences were not sensitive to the choice of the prior parameters. We sampled from the posterior using a Gibbs sampler.

Appendix B: Bayesian estimation for the GPD distribution

We describe the GPD model and then detail the MCMC algorithm used for model fitting. Bayesian inference for the GPD has been previously detailed in, for instance, Bermudez et al. (2001) and Diebolt et al. (2005). We tailor our MCMC algorithm to err on the side of conservatism in percentile estimates in the absence of ample data.

Model

GPD likelihood. After censoring the observed data X at a threshold u, we model $Y = X - u|X > u$ using a generalized Pareto distribution. The likelihood for the GPD distribution is :

$$\begin{aligned} l(\beta, \xi|y) &= \prod_{i=1}^{n} l(\beta, \xi|y_i) \text{ where} \\ l(\beta, \xi|y_i) &= \begin{cases} \frac{1}{\beta}(\frac{\xi y_i}{\beta} + 1)^{-\frac{\xi+1}{\xi}} & \text{if } \xi \neq 0 \\ \frac{1}{\beta}e^{-y_i/\beta}, & \text{if } \xi = 0 \end{cases} \end{aligned} \tag{6.3}$$

The parameter β is a scale parameter whereas ξ is a shape parameter that dictates the heaviness in the tails. Asymptotically, the GPD model is a limiting case for many known probability distributions. Specifically, the case $\xi < 0$ corresponds to short tailed distributions, such as the uniform; $\xi = 0$ includes exponentially shaped right tails and encompasses many common distributions such as the normal, lognormal, Weibull, and gamma distributions; $\xi > 0$ includes heavier tailed distributions, such as the Pareto and Student's t- distributions (Gomes and Guillou, 2015). Implementing an MCMC sampler for the GPD distribution is complicated by the sign of ξ, which determines the heaviness of the distribution tails. Bermudez et al. (2001) recommend simply choosing the sign of ξ based on maximum likelihood inferences for ξ. Herein, we restrict to $\xi \geq 0$, given that the validation metrics are targeted toward protecting against anti-conservatism due to heavy-tailed distributions.

Mixture representation. We consider the cases $\xi = 0$ and $\xi > 0$ as two separate models, $\mathscr{M}_{\xi=0}$ and $\mathscr{M}_{\xi>0}$. When $\xi = 0$, we parameterize the GPD distribution using β_0; when $\xi > 0$, we use parameter notation ξ, β.

$$\begin{aligned}
Y|\xi,\beta,\mathscr{M}_{\xi>0} &\sim GPD(\xi,\beta) \\
Y|\beta_0,\mathscr{M}_{\xi=0} &\sim GPD(0,\beta_0) \\
\beta|\mathscr{M}_{\xi>0} &\sim Unif(a_\beta,b_\beta) \\
\xi|\mathscr{M}_{\xi>0} &\sim Unif(a_\xi,b_\xi) \\
\beta_0|\mathscr{M}_{\xi=0} &\sim Unif(a_\beta,b_\beta) \\
\mathscr{M}_{\xi=0} &\sim Ber(w)
\end{aligned} \tag{6.4}$$

Prior distribution. We place equal prior weight on the cases $\xi = 0$ and $\xi > 0$, assuming $w = .5$. Parameters of the GPD distribution are often difficult to estimate (Bermudez et al., 2001); therefore, we use proper uniform, but disperse, empirical Bayes priors on β, β_0 and ξ to achieve good MCMC convergence.

We specify a uniform prior on β and β_0, namely $\beta, \beta_0 \sim U(a,b)$. We use an empirical Bayesian method to select a and b. Specifically, we calculate the maximum likelihood estimate and corresponding standard error for β, denoted $\hat{\beta}$ and $\hat{se}(\beta)$, and choose $\hat{\beta} \pm 15 * \hat{se}(\beta)$ for a,b (but truncating the minimum on a at 0). Maximum likelihood estimation was implemented using the `evir` `R` package (Pfaff and McNeil, 2012). We also considered an inverse-gamma prior on β, β_0, following Bermudez et al. (2001), selecting the hyperparameters again using an empircal Bayesian approach as well as using 'non-informative' hyperparameters of .01, .01. Results were similar using the uniform and inverse-gamma priors; all results in this paper use the uniform prior. For applications where β is small, the inverse gamma prior can perform poorly (Gelman et al., 2006), and we recommend rescaling or using the uniform prior in this case.

Posterior distribution and sampling. The posterior is a mixture of the two different models, $\mathscr{M}_{\xi>0}$ and $\mathscr{M}_{\xi=0}$. We follow the MCMC approach in Stephenson and Tawn (2004), using reverse-jump MCMC estimate the probabilities associated with each model. The MCMC algorithm is:

1. Calculate the maximum likelihood estimates of ξ and β as starting values.
2. Use Metropolis-within-Gibbs to estimate the posterior distribution each model $\mathscr{M}_{\xi=0}$ and $\mathscr{M}_{\xi>0}$ separately. For instance, for $\mathscr{M}_{\xi>0}$, the algorithm is:
   - Sample $\beta^{(t)}$ from its conditional posterior, $p(\beta|\xi^{(t-1)})$, using Metropolis-Hastings. For the proposal density, we use a normal proposal density centered at $\beta^{(t-1)}$ with the standard deviation selected to achieve good mixing.
   - Sample $\xi^{(t)}$ from its conditional posterior, $p(\xi|\beta^{(t)})$, using Metropolis-Hastings. We use a normal proposal density centered at $\xi^{(t-1)})$ with the standard deviation selected to achieve good mixing.
3. Use the reverse-jump MCMC algorithm described in Stephenson and Tawn (2004) to estimate $w = P(\mathscr{M}_{\xi=0}|y)$.

Tuning parameters were chosen so that the acceptance rate is between 0.4 and 0.6 on average. For the example problems in Section 3.2, we ran 50,000 MCMC iterations, with 10,000 runs discarded as burn-in (fewer runs were used for the simulation studies). We evaluated model convergence by running multiple chains and examining traceplots.

Simulation study

We conducted a simulation study to evaluate the performance of the Bayesian estimation procedure relative to frequentist inference on the GPD model. Specifically, as a frequentist alternative, we consider bootstrapping the maximum likelihood estimates for the GPD parameters as in, for instance, (Diebolt et al., 2007). We used a simple parametric bootstrap for uncertainty quantification, with confidence intervals constructed using the basic bootstrap method (Davison and Hinkley, 1997). Parametric bootstrapping simply requires constructing bootstrap samples by re-sampling data from the fitted GPD model using MLE and then re-fitting the GPD model to all of the bootstrap samples.

We consider the following cases in the simulation:

- Data are simulated from the standard normal and t_5 distributions.
- We estimate the following percentiles: .99, .995, .999.
- We simulate data with sample sizes $n = 100, 250$, and 1000.
- We choose u such that we keep the largest 10% of the simulated data.

For each case, we run the MLE method and the Bayesian method a total of 500 times. For the Bayesian method, we produce a chain of 1200 values using a burn-in of 200 (running many fewer

MCMC samples than the final results due to the computational expense). We examined one-sided 95% confidence interval coverage (to make sure that the method effectively bounds the true percentile 95% of the time), as well as bias in the percentile estimate, measured using the true versus median estimate over the simulations. We used the median estimate in simulation because, when data were sparse, the distribution of GPD estimates was highly right skewed (as desired, due to the lack of data for extreme percentile estimation).

Based on the simulation results (Table 6.1 and Figure 6.2), we conclude that the Bayesian model errs on the side of conservatism, while the frequentist method can be anti-conservative. The errors in percentile estimation were closely related to how much data was available relative to both the return-level and the non-parametric sample size requirement (Equation 3.2). In general, the Bayesian method provides higher coverage than the frequentist method, and tends to be closer to the nominal 95% coverage (Figure 6.2). The frequentist method tends to be anti-conservative, with coverage nearly always well below the desired 95% When the sample size is small the Bayesian method tends to be heavily conservative, yielding higher coverage than the maximum likelihood method.

When the true value of the percentile of interest is smaller than u, both methods fail (as expected), and thus we emphasize the importance of choosing u with care. Specifically, we can check that the fraction of the data used for tail estimation, denoted f_u, satisfies the condition:

$$f_u \geq 1 - F_B^{-1}(\alpha | n, p)/n \tag{6.5}$$

Here, $F_B^{-1}(\cdot | n, p)$ is the Binomial quantile function with parameters n and p. If this condition holds, then at least one uncensored data point will be less extreme than the quantile of interest with probability no less than $1 - \alpha$. For instance, if $Q_{0.98}$ is the quantile of interest and there are $n = 100$ observations, then we note that:

$$1 - F_B^{-1}(0.01, 100, 0.98)/100 = 0.06 \tag{6.6}$$

Thus for the choice $\alpha = 0.01$, we should choose u so that f_u, the fraction of data used for tail estimation, is no smaller than 0.06. We can make the choice $f_u = 0.1$ and conclude that at least one of the uncensored data points will be smaller than $Q_{0.98}$ with probability greater than 0.99.

Typically, this condition will be less restrictive when the quantile of interest is very extreme. For illustration, repeating the above example for $Q_{0.999}$ and $Q_{0.9999}$ gives constraints $f_u \geq 0.01$ and $f_u \geq 0$ respectively.

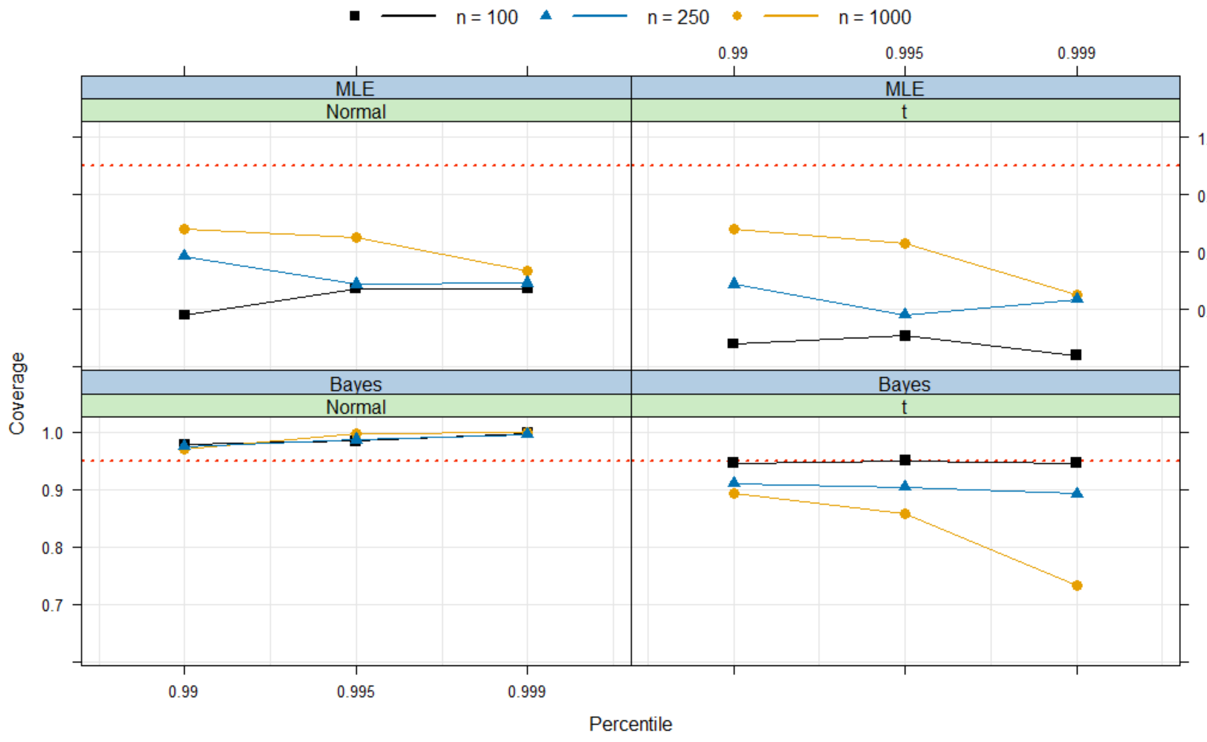

Figure 6.2: Coverages from the simulation study plotted versus percentile for the maximum likelihood (top panels) and Bayesian (bottom panels) methods, for data generated from the standard normal (left panels) and the t-distribution with 5 degrees of freedom (right panels). The dotted line is drawn at 0.95 representing desired coverage.

Distr.	p	n	Truth	Median (MLE)	Median (Bayes)
Normal	100	0.99	2.33	2.28	2.50
Normal	250	0.99	2.33	2.30	2.40
Normal	1000	0.99	2.33	2.33	2.39
Normal	100	0.995	2.58	2.52	2.88
Normal	250	0.995	2.58	2.54	2.75
Normal	1000	0.995	2.58	2.57	2.71
Normal	100	0.999	3.09	3.03	3.97
Normal	250	0.999	3.09	3.04	3.57
Normal	1000	0.999	3.09	3.03	3.48
t_5	100	0.99	3.36	3.25	3.76
t_5	250	0.99	3.36	3.31	3.45
t_5	1000	0.99	3.36	3.38	3.42
t_5	100	0.995	4.03	3.86	4.54
t_5	250	0.995	4.03	3.92	4.13
t_5	1000	0.995	4.03	4.05	4.08
t_5	100	0.999	5.89	5.55	57.27
t_5	250	0.999	5.89	5.74	6.11
t_5	1000	0.999	5.89	5.75	5.69

Table 6.1: Percentile estimates using Bayesian and frequentist inference. The true percentile is compared to the median across 500 simulations.

References

AIAA. "Guide for the Verification and Validation of Computational Fluid Dynamics Simulations, American Institute of Aeronautics and Astronautics." Technical report (1998). AIAA-G-077-1998.

Aitchison, J. "Two Papers on the Comparison of Bayesian and Frequentist Approaches to Statistical Problems of Prediction: Bayesian Tolerance Regions." *Journal of the Royal Society, Series B*, 26(2):161–175 (1964).

Aldor-Noiman, S., Brown, L., Buja, A., Rolke, W., and Stine, R. "The Power to See: A New Graphical Test of Normality." *Amer. Stat.*, 67:249–260 (2013).

Atwood, C., LaChance, J., Martz, H., Anderson, D., Englehardt, M., Whitehead, D., and Wheeler, T. "Handbook of parameter estimation for probabilistic risk assessment." *US Nuclear Regulatory Commission, Washington, DC, Report No. NUREG/CR-6823* (2003).

Barker, R. J. and Link, W. A. "Bayesian multimodel inference by RJMCMC: A Gibbs sampling approach." *The American Statistician*, 67(3):150–156 (2013).

Bergquist, M. L. *Caution Using Bootstrap Tolerance Limits with Application to Dissolution Specification Limits*. ProQuest (2006). Doctoral dissertation.

Bermudez, P. d. Z., Turkman, M. A., and Turkman, K. "A predictive approach to tail probability estimation." *Extremes*, 4(4):295–314 (2001).

Chivers, C. *MHadaptive: General Markov Chain Monte Carlo for Bayesian Inference using adaptive Metropolis-Hastings sampling* (2012). R package version 1.1-8.
URL `https://CRAN.R-project.org/package=MHadaptive`

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. *An introduction to statistical modeling of extreme values*, volume 208. Springer (2001).

Csörgő, S. and Faraway, J. "The exact and asymptotic distributions of Cramér-von Mises Statistics." *J. R. Statist. Soc. B.*, 58:221–234 (1996).

D'Agnostino, R. and Stephens, M. (eds.). *Goodness-of-fit Techniques*, volume 68 of *Statistics: a Series of Textbooks and Monographs*. New York: Marcel Dekker, Inc., 1st edition (1986).

Davison, A. C. and Hinkley, D. V. *Bootstrap methods and their application*, volume 1. Cambridge university press (1997).

De Haan, L. and Ferreira, A. *Extreme value theory: an introduction*. Springer Science & Business Media (2007).

Diebolt, J., El-Aroui, M.-A., Garrido, M., and Girard, S. "Quasi-conjugate Bayes estimates for GPD parameters and application to heavy tails modelling." *Extremes*, 8(1-2):57–78 (2005).

Diebolt, J., Garrido, M., and Girard, S. "A goodness-of-fit test for the distribution tail." *Topics in Extreme Values*, 95–109 (2007).

Diegert, K., Klenke, S., Novotny, G., Paulsen, R., Pilch, M., and Trucano, T. "Toward a More Rigorous Application of Margins and Uncertainties within the Nuclear Weapons Life Cycle – A Sandia Perspective." Technical report (2007). SAND2007-6219.

Eardley, D. "Quantifications of Margins and Uncertainties." Technical report (2005).

Efron, B. and Tibshirani, R. J. *An introduction to the bootstrap*. CRC press (1994).

English, J. and Taylor, G. "Process capability analysis—a robustness study." *The international journal of production research*, 31(7):1621–1635 (1993).

Fernholz, L. T. and Gillespie, J. A. "Content-corrected tolerance limits based on the bootstrap." *Technometrics*, 43(2):147–155 (2001).

Ferson, S., Oberkampf, W., and Ginzburg, L. "Model validation and predictive capability for the thermal challene problem." *Computer Methods in Applied Mechanics and Engineering*, 197(29–32):2408–2430 (2008).

Fisher, K. "Statistical tests." *Nature*, 136:474 (1935).

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA (2014).

Gelman, A. et al. "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian analysis*, 1(3):515–534 (2006).

Gomes, M. I. and Guillou, A. "Extreme value theory and statistics of univariate extremes: A review." *International Statistical Review*, 83(2):263–292 (2015).

Greenland, S. and Poole, C. "Living with p values: resurrecting a Bayesian perspective on frequentist statistics." *Epidemiology*, 24(1):62–68 (2013).

Hahn, G. and Meeker, W. "Pitfalls and practical considerations in product life analysis, part 1: Basic concepts and dangers of extrapolation." *Journal of Quality Technology*, 14(3):144–152 (1982).

Haimes, Y. Y. and Lambert, J. H. "When and How Can You Specify a Probability Distribution When You Don't Know Much? II." *Risk Analysis*, 19(1):43–46 (1999).

Hamada, M., Johnson, V., Moore, L., and Wendelberger, J. "Bayesian Prediction Intervals and Their Relationship to Tolerance Intervals." *Technometrics*, 46:452–459 (2004).

Helton, J. C. "Conceptual and Computational Basis for the Quantification of Margins and Uncertainty." Technical report (2009). SAND2009-3055.

—. "Quantification of margins and uncertainties: Conceptual and computational basis." *Reliability Engineering & System Safety*, 96(9):976–1013 (2011).

Ho, Y. H. and Lee, S. M. "Calibrated interpolated confidence intervals for population quantiles." *Biometrika*, 92(1):234–241 (2005).

Ho, Y. H., Lee, S. M., et al. "Iterated smoothed bootstrap confidence intervals for population quantiles." *The Annals of statistics*, 33(1):437–462 (2005).

Horn, P. S. "Quasi-nonparametric upper tolerance regions based on the bootstrap." *Communications in Statistics-Theory and Methods*, 21(12):3351–3367 (1992).

Hughey, R. L. "A survey and comparison of methods for estimating extreme right tail-area quantiles." *Communications in Statistics-Theory and Methods*, 20(4):1463–1496 (1991).

Hutson, A. D. "Calculating nonparametric confidence intervals for quantiles using fractional order statistics." *Journal of Applied Statistics*, 26(3):343–353 (1999).

Janiga, I. and Garaj, I. "One-sided Tolerance Factors of Normal Distributions with unknown mean and variability." *Measurement Science Review*, 6:12–16 (2006).

Juárez, M. A. and Steel, M. F. J. "Model-Based Clustering of Non-Gaussian Panel Data Based on Skew-t Distributions." *Journal of Business & Economic Statistics*, 28(1):52–66 (2010).

Kerns, G. *Introduction to Probability and Statistics Using R*. G. Jay Kerns, 1st edition (2010).

Khorsandi, J. and Aven, T. "Incorporating assumption deviation risk in quantitative risk assessments: A semi-quantitative approach." *Reliability Engineering & System Safety*, 163:22–32 (2017).

King, G., Tomz, M., and Wittenberg, J. "Making the most of statistical analyses: Improving interpretation and presentation." *American journal of political science*, 347–361 (2000).

Krishnamoorthy, K. and Mathew, T. *Statistical Tolerance Regions: Theory, Applications, and Computation*. Wiley Series in Probability and Statistics. Hoboken, N.J, USA: John Wiley & Sons, Inc. (2009).

Lanes, C. E. and Hill, R. G. "(U) Supplemental requirements, Quantification of Margin and Uncertainties, B61-12." (2016).

Loy, A., Follet, L., and Hofmann, H. "Variations of Q-Q Plots: The Power of Our Eyes!" *Amer. Stat.*, 70:202–214 (2015).

Mahadevan, S. and Rebba, R. "Validation of reliability computational models using Bayes networks." *Reliability Engineering & System Safety*, 87(2):223–232 (2005).

Meeker, W. Q. and Escobar, L. A. *Statistical methods for reliability data*. John Wiley & Sons (2014).

Michael, J. and Schucany, W. "Analysis of data from censored samples." In D'Agnostino, R. and Stephens, D. (eds.), *Goodness-of-fit Techniques*, chapter 11. New York: Dekker (1986).

Miller, R. *Survival Analysis*. New York: John Wiley & Sons, Inc. (1981).

Montgomery, D. C., Runger, G. C., and Hubele, N. F. *Engineering statistics*. John Wiley & Sons (2009).

Moon, Y.-I., Lall, U., and Bosworth, K. "A comparison of tail probability estimators for flood frequency analysis." *Journal of Hydrology*, 151(2-4):343–363 (1993).

Mosleh, A. "Hidden sources of uncertainty: judgment in the collection and analysis of data." *Nuclear Engineering and Design*, 93(2-3):187–198 (1986).

Mullins, J., Ling, Y., Mahadevan, S., Sun, L., and Strachan, A. "Separation of aleatory and epistemic uncertainty in probabilistic model validation." *Reliability Engineering & System Safety*, 147:49–59 (2016).

Mullins, J. and Mahadevan, S. "Bayesian Uncertainty Integration for Model Calibration, Validation, and Prediction." *J. Verif. Valid. Uncert.*, 1 (2016).

National Research Council. *Evaluation of Quantification of Margins and Uncertainties Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile*. National Academies Press (2008).

Newcomer, J., Rutherford, B., Thomas, E., Bierbaum, R., Hickman, L., Lane, J., Fitchett, S., Urbina, A., Robertson, A., and Swiler, L. "Handbook of Statistical Methodologies for QMU." Technical report (2012). SAND2012-9603.

Newcomer, J. T. "A new approach to quantification of margins and uncertainties for physical simulation data." *SAND2012-7912* (2012).

NIST/SEMATECH. "e-Handbook of Statistical Methods." (????).
URL http://www.itl.nist.gov/div898/handbook

Oberkampf, W. L. and Barone, M. F. "Measures of agreement between computation and experiment: validation metrics." *Journal of Computational Physics*, 217(1):5–36 (2006).

Pfaff, B. and McNeil, A. *evir: Extreme Values in R* (2012). R package version 1.7-3.
URL https://CRAN.R-project.org/package=evir

Pilch, M., Trucano, T. G., and Helton, J. C. "Ideas underlying quantification of margins and uncertainties (QMU): a white paper." *Unlimited Release SAND2006-5001, Sandia National Laboratory, Albuquerque, New Mexico*, 87185:2 (2006).

—. "Ideas underlying the quantification of margins and uncertainties." *Reliability Engineering & System Safety*, 96(9):965–975 (2011).

Pradlwarter, H. and Schuëller, G. “The use of kernel densities and confidence intervals to cope with insufficient data in validation experiments.” *Computer Methods in Applied Mechanics and Engineering*, 197(29):2550–2560 (2008).

Rahman, M. and Govindarajulu, Z. “A modification of the test of Shapiro and Wilk for normality.” *J. Appl. Stat.*, 24:219–236 (1997).

Razali, N. and Wah, Y. “Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests.” *J. Stat. Mod. Anal.*, 2:21–33 (2011).

Razali, N. M., Wah, Y. B., et al. “Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests.” *Journal of statistical modeling and analytics*, 2(1):21–33 (2011).

Rebba, R. and Mahadevan, S. “Computational methods for model reliability assessment.” *Reliability Engineering & System Safety*, 93(8):1197–1207 (2008).

Roeder, K. and Wasserman, L. “Practical Bayesian density estimation using mixtures of normals.” *Journal of the American Statistical Association*, 92(439):894–902 (1997).

Romero, V., Swiler, L., Urbina, A., and Mullins, J. “A Comparison of Methods for Representing Sparsely Sampled Random Quantities.” (2013).

Schenker, N. “Qualms about bootstrap confidence intervals.” *Journal of the American Statistical Association*, 80(390):360–361 (1985).

Scholz, F. “Nonparametric tail extrapolation.” *Boeing Information & Support Services ISSTECH-95-014, Seattle, WA* (2005).

—. “The bootstrap small sample properties.” *Boeing Computer Services, Research and Technology, Tech. Rep* (2007).

Segalman, D. J., Paez, T. L., and Bauman, L. E. “A Robust Approach to Quantification of Margin and Uncertainty.” *Journal of Verification, Validation and Uncertainty Quantification*, 2 (2017).

Sharp, D., Wallstrom, T., and Wood-Schulz, M. “Physics package confidence: “one” vs. “1.0”.” *Proceedings of the NEDPC 2003* (2003).

Song, E., Nelson, B. L., and Pegden, C. D. “Advanced tutorial: Input uncertainty quantification.” In *Simulation Conference (WSC), 2014 Winter*, 162–176. IEEE (2014).

Spiegelhalter, D., Pearson, M., and Short, I. “Visualizing uncertainty about the future.” *science*, 333(6048):1393–1400 (2011).

Stan Development Team. “Pystan: the Python interface to Stan.” (2016 Version 2.14.0.0). URL `http://mc-stan.org`

Stephens, M. A. “EDF statistics for goodness of fit and some comparisons.” *Journal of the American statistical Association*, 69(347):730–737 (1974).

Stephenson, A. and Tawn, J. "Bayesian inference for extremes: accounting for the three extremal types." *Extremes*, 7(4):291–307 (2004).

Su, S. "Confidence intervals for quantiles using generalized lambda distributions." *Computational Statistics & Data Analysis*, 53(9):3324–3333 (2009).

Thomas, E. V. "A Statistical Perspective on Highly Accelerated Testing." Technical report (2015). SAND2015-0927.

Urbina, A., Mahadevan, S., and Paez, T. L. "A Bayes network approach to uncertainty quantification in hierarchically developed computational models." *International Journal for Uncertainty Quantification*, 2(2) (2012).

Villaseñor-Alva, J. A. and González-Estrada, E. "A bootstrap goodness of fit test for the Generalized Pareto Distribution." *Computational Statistics & Data Analysis*, 53(11):3835–3841 (2009).

Wakefield, J. *Bayesian and frequentist regression methods*. Springer Science & Business Media (2013).

Walker, E. and Nowacki, A. S. "Understanding equivalence and noninferiority testing." *Journal of general internal medicine*, 26(2):192–196 (2011).

Waller, L. A. and Turnbull, B. W. "Probability plotting with censored data." *The American Statistician*, 46(1):5–12 (1992).

Wallstrom, T. C. "Quantification of margins and uncertainties: A probabilistic framework." *Reliability Engineering & System Safety*, 96(9):1053–1062 (2011).

Wilks, S. S. "Determination of sample sizes for setting tolerance limits." *The Annals of Mathematical Statistics*, 12(1):91–96 (1941).

Xie, M.-g. and Singh, K. "Confidence distribution, the frequentist distribution estimator of a parameter: a review." *International Statistical Review*, 81(1):3–39 (2013).

Yeo, I.-K. and Johnson, R. A. "A new family of power transformations to improve normality or symmetry." *Biometrika*, 954–959 (2000).

Young, D. S. and Mathew, T. "Improved nonparametric tolerance intervals based on interpolated and extrapolated order statistics." *Journal of Nonparametric Statistics*, 26(3):415–432 (2014).

Zouaoui, F. and Wilson, J. R. "Accounting for parameter uncertainty in simulation input modeling." *Iie Transactions*, 35(9):781–792 (2003).

Sandia National Laboratories